Performance Analysis & Traffic Modelling in Broadband Integrated Services Digital Networks

by

Daryoush Habibi B.E.(Hons.)

Department of Electrical and Electronic Engineering

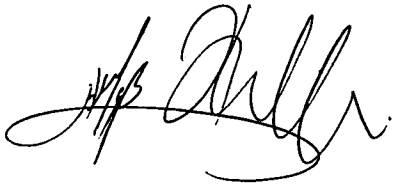
Submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

University of Tasmania

October 1993

Statement of Originality

This thesis contains no material which has been accepted for the award of any other degree or diploma in any tertiary institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person, except when due reference is made in the text.



Daryoush Habibi

To my parents.

Abstract

Broadband Integrated Services Digital Networks (B-ISDN) will provide the ability to support a wide range of services using Asynchronous Transfer Mode (ATM) as the transfer technique. While ATM provides the flexibility to integrate a large number of services economically, the challenge to teletraffic engineers is to ensure that an acceptable grade of service is provided for each type of traffic. Research in network performance and traffic modelling is required to develop network operating strategies which will meet these goals.

This thesis is mainly concerned with performance analysis models for B-ISDN. In performance modelling of access nodes, several strategies are studied for mixing loss-sensitive traffic with delay-sensitive traffic in both TDM and ATM environments. The performance of these strategies are analysed by different methods. Next, the problem of statistical multiplexing of interactive data, interactive images and variable bit rate (VBR) video traffic in an ATM access node is considered. In this situation, the effect of link rate to source rate ratio and also the effect of priority encoding of the VBR video on the performance of the ATM access node are studied. Next, a strategy is proposed for statistically multiplexing a range of constant bit rate services with an aggregate of variable bit rate services at an ATM access node. Performance parameters for both service types are evaluated by analysis and also by simulation.

Accurate source models are an essential component of any access node problem. The traffic generated from video services will greatly influence the overall performance and data requirements in B-ISDN, and the next area considered in this thesis is modelling of video traffic. Several video models in the literature

are reviewed and a few models based on the concept of hidden Markov models are examined for modelling variable bit rate video traffic. Network performance based on these models is investigated.

Another area that is covered in this thesis is performance modelling for those parts of the network that are subject to traffic with periodically varying rates. A computational probability analysis is presented for queues with cyclo-stationary arrivals and/or cyclo-stationary service rates. Such traffic patterns may arise in a variety of telecommunication and computer networks, including B-ISDN.

Acknowledgements

Firstly I would like to thank my supervisors Doctor David J.H. Lewis and Professor D. Thong Nguyen. In particular I am very grateful to Doctor David Lewis for being a valuable friend, for his thoughtful ideas that have helped me greatly in my research, and for his guidance and support throughout these studies. I am also very grateful to Professor Thong Nguyen for persuading me to attempt this higher degree, for initiating a research contract which assisted me financially while undertaking this degree, and for his support and encouragement during my studies. I would also like to thank all the staff of the Electrical & Electronic Engineering Department at the University of Tasmania, especially Mr Gregory Thé, Mr Peter Watt, Dr Richard Langman and Dr Richard Lane for their kind treatment and friendship which made these years very comfortable and enjoyable. Many thanks go to my fellow graduate students who made the study environment cooperative and pleasant, particularly Mark Stoksik and Andrew Bainbridge-Smith who went out of their way to help with the proof reading of this manuscript.

At last and not least, I would like to thank my beloved wife, Vida Ghoddousi, for keeping up with me and for being very caring and supportive throughout the course of my studies.

Contents

List of Figures

List of Tables

List of Acronyms

AAL	ATM adaptation layer
ATM	Asynchronous transfer mode
B-ISDN	Broadband integrated services digital network
CBR	Constant bit rate
CCITT	International Telegraph and Telephone Consultative Committee
CLP	Cell loss priority
CS	Convergence sublayer
CSMA/CD	Carrier sense multiple access with collision detection
DQDB	Distributed queue dual bus
FDDI	Fibre distributed data interface
FDM	Frequency division multiplexing
FIFO	First in first out
GFC	Generic flow control
HEC	Header error control
HDTV	High definition television
IDN	Integrated digital network
ISDN	Integrated services digital network
IU	Interworking unit
Kbps	Kilo bits per second
LAN	Local area network
LSI	Large scale integration
MAC	Medium access control
MAN	Metropolitan area network
Mbps	Mega bits per second
NNI	Network-node interface

NT1	Network termination type 1
NT2	Network termination type 2
OSI	Open system interconnections
PCM	Pulse code modulation
PDU	Protocol data unit
PM	Physical medium
PRM	Protocol reference model
PSTN	Public switched telephone network
PT	Payload type
QOS	Quality of service
SAP	Service access point
SAR	Segmentation and reassembly
TA	Terminal adaptor
TC	Transmission convergence
TDM	Time division multiplexing
TE	Terminal equipent
TE1	Terminal equipment type 1
TE2	Terminal equipment type 2
UNI	User-network interface
VBR	Variable bit rate
VC	Virtual channel
VCC	Virtual channel connection
VCI	Virtual channel identifier
VLSI	Very large scale integration
VP	Virtual path
VPC	Virtual path connection
VPI	Virtual path identifier

Preface

The original purpose of this research was to study performance analysis models for Broadband Integrated Services Digital Networks (B-ISDNs), especially those related to connection admission control. During the course of these studies it was realised that accurate traffic models are essential in performance studies of high speed networks. In particular it is important to model video services accurately, as the traffic generated from these services will greatly influence the overall performance and data requirements in broadband networks. Two chapters of this thesis deal with video traffic modelling, providing a literature survey of traffic models for video services and investigating the performance of hidden Markov models for modelling VBR video traffic. Another extension to this work is the development of performance analysis models for the queueing systems that have periodic variations in the arrival rate and/or service rate of their traffic.

The work presented in this thesis was initiated as part of a research contract between Telecom Australia Research Laboratories and the Department of Electrical & Electronic Engineering, University of Tasmania. The research for this thesis took place at the University of Tasmania between April 1990 and October 1993.

Thesis Organisation

This thesis is organised into nine chapters. The first chapter gives an introduction to the evolution of B-ISDNs, provides a general description of B-ISDNs, and outlines the problems associated with traffic control and resource management in broadband networks. Chapter 2 outlines and analyses two call access control strategies that may be suitable for networks carrying a mixture of two types

of traffic, narrowband (NB) and wideband (WB). The NB traffic and the WB traffic may be identified as data traffic and fixed bit rate video traffic respectively in a broadband network. Several methods are used for the analysis of these strategies. Chapter 3 uses simulation tools to study the problem of multiplexing interactive data, interactive images, and video traffic in an ATM access node and investigates the effect of reducing the ratio of the link bit rate to peak bit rates of the incoming traffic on the performance of the access node. It also investigates the effect of priority encoding of the video traffic on the performance of the access node. Chapter 4 outlines and analyses a strategy for an access node where a range of constant bit rate (CBR) and variable bit rate (VBR) services are multiplexed. Chapter 5 provides a literature survey of traffic models for video services. Chapter 6 continues the study of video traffic modelling and investigates the performance of hidden Markov models for VBR video services. The correlation studies of the traffic generated by a VBR video codec indicates the presence of periodic variations in the cell stream generated by VBR video services. This sets the scene for Chapters 7 and 8 which present an analysis method for queues that have periodic variations in the arrival rate and/or service rate of their traffic. Finally, Chapter 9 provides a summary of the contributions and the major results of this thesis and suggests some extensions to this work for future research.

Supporting Publications

This research resulted in six technical reports to Telecom Australia Research Laboratories (four reports on Telecom Research Laboratories contract no. 7174 and two reports on Telecom Research Laboratories contract no. 7332), and seven journal and conference publications as listed below:

1. Daryoush Habibi & DJH Lewis, *' A Solution for Cyclo-Stationary Queueing Systems'*, Submitted to IEE Electronics Letters.

2. Daryoush Habibi, DJH Lewis, DT Nguyen & Jason Pieloor, *'Analysis of an Access Node Multiplexer in a System Serving CBR and VBR Traffic'*, to appear in *'Computer Communications'*, Volume 16, Number 12, December

1993.

3. Daryoush Habibi & DJH Lewis, *'Queues with Periodic Input and Output Rates'*, Proceedings of The Australian Broadband Switching & Services Symposium '93, pages 225-233, Wollongong, July 1993.

4. Daryoush Habibi, *'A Hidden Markov Model for Modelling VBR Video'*, Proceedings of The Seventh Australian Teletraffic Research Seminar, pages 181-190, Adelaide, November 1992.

5. Daryoush Habibi, DJH Lewis, DT Nguyen & Jason Pieloor, *'Performance of A Multiplexer in a B-ISDN Network with STM and ATM Traffic'*, Proceedings of The Australian Broadband Switching & Services Symposium '92, pages 691-698, Melbourne, July 1992.

6. Daryoush Habibi, DJH Lewis & DT Nguyen, *'Access Control in ATM Networks Carrying Video, Data and Interactive Images'*, Proceedings of The Australian Broadband Switching & Services Symposium '91, pages 165-173, Sydney, July 1991.

7. DJH Lewis & Daryoush Habibi, *Analysis vs. Simulation: The Computational Effort'*, Proceedings of The Australian Broadband Switching & Services Symposium '91, Sydney, July 1991.

Reprints of publications 2, 3 and 4 are provided in the section titled 'Reprints of Selected Papers' at the end of this thesis.

Chapter 1

Introduction

1.1 Introduction

Telecommunications networks have gone through many changes since the invention of Morse code in the late 1830s [1]. The early communication networks, telegraph networks, were conceptually digital (morse code). Then, there was the migration towards analog systems (analog telephony) which dominated the telecommunications industry for a few decades. In the recent decades these analog systems are being phased out to be replaced by the more efficient, higher quality digital networks. Since the 1830's many telecommunication services have entered the scene: telephone, telex, television, facsimile, electronic mail, etc. and the list is growing very fast. These new services have revolutionised lifestyles by introducing new means of communication and learning and by facilitating many services that are used in our daily lives such as banking, travel, shopping, etc. Today, there are many telecommunications networks, heterogeneous in nature and service specific [2]. With the evolution of these specialised networks, the ultimate goal in telecommunications has become network flexibility and service independence. The latest development in telecommunication is the concept of the Broadband Integrated Services Digital Network (B-ISDN). B-ISDN my be considered as the most revolutionary network concept in the recent decades. It will deliver many high bit rate services including a wide range of video and image services that will have a big impact on the commercial and residential customers.

In section 1.2, the earlier telecommunication technologies are briefly outlined. Section 1.3 outlines the service categories that are expected in a mature B-ISDN. Section 1.4 describes a protocol reference model for B-ISDN which reflects the principles of layered communication of the reference model of open system interconnections. Section 1.5 discusses the concept of Asynchronous Transfer Mode in relation to B-ISDN. Section 1.6 provides a review of traffic control and resource management in B-ISDN. Finally, section 1.7 describes the major thrusts of this thesis and lists the areas to which this thesis makes contributions.

1.2 The Evolution of Broadband Networks

The purpose of this section is to give a summary of the network technologies that step-by-step have evolved into the introduction of B-ISDNs. The topics covered in this section are circuit switching, packet switching, Integrated Digital Networks, Local Area Networks, Metropolitan Area Networks, integrated Services Digital Networks, and finally Broadband Integrated Services Digital Networks.

1.2.1 Circuit Switching

Circuit switching was originally designed for voice traffic and it is still the dominant technology for voice communication and a major alternative for data communication. As the name implies, circuit switching is based on establishing a dedicated path between the calling and the called parties before any information is transmitted. This means that a dedicated capacity is maintained on all the links through which the call has been routed for the duration of the call, regardless of whether information is being transferred on the path. Let us briefly look at the fundamental processes in a circuit switched network by looking at the network diagram shown in Figure 1.1. Let us assume that subscriber B wants to establish a connection to communicate with subscriber D. This process consists of three distinct phases as described below:

- *Connection phase:* In this phase subscriber B sends a message to his server node (node 1) requesting a connection to subscriber D. Usually there is a dedicated line between B and node 1. Node 1 selects the outgoing link to

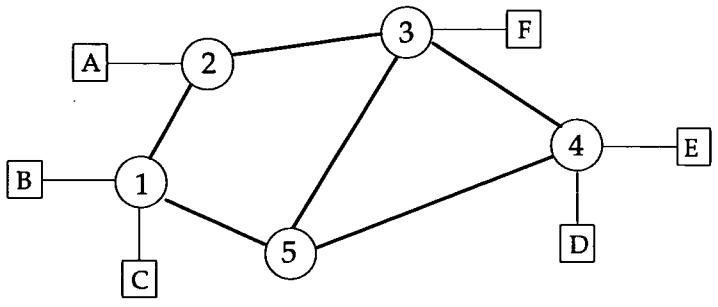


Figure 1.1: A switched network

node 5 (based on the routing information available to it and based on other parameters such as the utilisation of the links and cost), allocates a channel on that link and passes on the connection request to node 5. Similarly node 5 allocates a channel on its link to node 3 and sends the connection request to node 3. The connection request is delivered to node 4 in a similar manner. Node 4 then checks to find if its line to subscriber D is busy or is willing to accept the connection request. On acceptance of the call by D, subscriber B is notified and the connection is completed.

- *Information transfer phase:* During this phase information can be transferred between the two parties in (usually) full duplex mode.

- *Disconnect phase:* Either of the parties can terminate the connection in which case a signal will be sent through to all of the nodes that maintain the connection to de-allocate the capacity dedicated to that connection.

Some main advantages of circuit switching are high reliability (e.g. low delay at each node and a fixed data transmission rate), relatively simple routing, and transparency. One of the main disadvantages of circuit switching is low efficiency for traffic that is bursty in nature, such as interactive data traffic. For example a terminal and a computer that are connected via a dedicated circuit will only use the circuit for occasional short bursts of traffic and the circuit will be idle for most of the time. Another disadvantage of circuit switching is that when two data terminal equipments are connected through a circuit switched network, they must transmit and receive at the same baud rate. This places an undesirable constraint on many data communication applications.

1.2.2 Packet Switching

The concept of packet switching started in early 70's with the aim of providing a fast, efficient, flexible and low cost method of data communications. The basic idea in packet switching is to divide the data into segments called packets (see Figure 1.2). Each packet has a header that contains the necessary information for routing of the packet through to the destination. The header may also contain extra control information, for example about the priority of the packet or about the maximum delay that it can tolerate. The network normally puts an upper limit on the size of the packet. Therefore if the data that the user wishes to send is larger than the maximum packet size, the data is broken into several packets and these packets are sent in sequence.

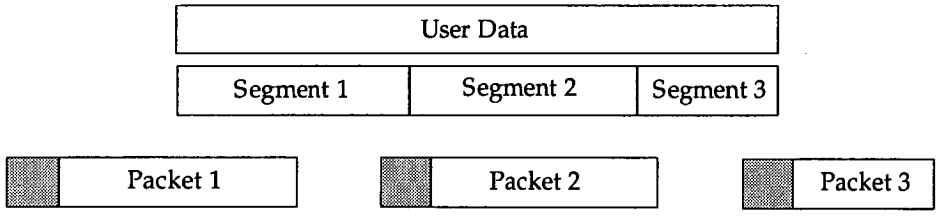


Figure 1.2: Packetising the data

There are two approaches that can be taken for routing the packets from source to destination in a packet switched network. These are known as *datagram* and *virtual circuit* methods and are briefly described here.

Datagram

There is no call setup phase in the datagram approach and each packet is sent independent of the previous packets. As mentioned earlier each packet has a header that contains information about the source and the destination of the packet. Packets are briefly buffered at each node. The nodes must process the information in the header and decide on the outgoing link on which the packet can be sent. This decision is usually based on the state of the nodes along alternative paths in the network. Hence as the queue size of various nodes varies with time, the path between the source and the destination can vary and as a result the

packets may arrive at the destination out of sequence. The Datagram approach is suitable for connectionless services. In cases where only a small amount of data is to be transferred, call setup time can be a major overhead on the network. This overhead is avoided in the datagram approach. Also network congestion is handled very well with this approach as each node will try to forward the packet via the least loaded node towards the destination.

Virtual Circuit

This approach has some similarities to circuit switching in that the entire path between the two users must be determined before any data packets are transferred. The difference of this approach to circuit switching is that the path is not dedicated. As an example let us reconsider the generic network shown in Figure 1.1 but now assume that it is a packet switched network. Again assume that subscriber B wants to communicate with subscriber D. B sends a *call request* packet to node 1. Node 1 can either choose node 2 or node 5 as the next node. If node 5 is selected, node 1 sends the call request packet and all the subsequent data packets to node 5. Node 5 selects node 4 and passes on the call request packet to it. Node 4 delivers the call request packet to D. If D decides to accept the connection, it will send a call accept packet that will be delivered back to B through nodes 4, 5 and 1 respectively. Data can then be transferred between B and D through the virtual path that has already been established. The connection can be terminated by either of the parties transmitting a *clear request* packet to the other party through the virtual circuit nodes.

The benefits of virtual circuit are: simplification of routing once a path has been established, and, allowing the components within the network to distinguish easily amongst different traffic flows for purposes of admission, congestion control, security, billing and so forth [3]. Because of its nature, the virtual circuit approach is suited well to connection oriented services that require error control and sequencing. Although there is some overhead due to the call setup time, for long transactions this will be outweighed by the savings in the processing time of each packet at the nodes along the virtual circuit, because there is no routing decision to make for each individual packet.

1.2.3 Integrated Digital Networks

The current telecommunication networks around the world are rapidly moving towards Integrated Digital Networks (IDN). The word *Integrated* in this context refers to the integration of transmission and switching and should not be mistaken with the integration of services. The major force behind the evolution of IDN has been the desire to achieve economical voice communications. With the advent of LSI and then VLSI the costs of producing digital equipment have declined significantly and the trend is continuing. The costs of manufacturing analog equipment have also dropped but not to the extent seen for digital equipment. In an IDN network all signals are treated as a stream of binary digits. Digital transmission and digital switching both require some processing of conventional analog signals such as voice before they can be transferred through IDN. In the first instance, the original analog signal (voice) is digitized to produce a stream of binary data. The digital data is then fed to a modem to produce an analog signal. There is a difference between this analog signal and the original analog voice signal in that the new analog signal is a representation of binary data and therefore digital transmission techniques can be applied to it.

The multiplexing technique used in an integrated digital network is *Time Division Multiplexing* (TDM). In TDM, time is divided into frames. Each time frame is divided into several time slots, and has a synchronisation bit. The time slots are allocated to various connections in a predetermined manner. Incoming links are sampled at appropriate frequency and samples are loaded into corresponding time slots. The sequence of the n^{th} time slots in consecutive time frames constitute the *nth* channel on the outgoing link. TDM in an integrated digital environment has cost and quality advantages over *Frequency Division Multiplexing* (FDM) in a non-integrated circuit switched network.

With digital transmission, the loss of quality resulting from switching nodes and relay points in the network is minimal compared to analog transmission. Also, IDN technology is significantly cheaper at switching nodes compared to its analog equivalent that uses FDM. To picture this, let us consider Figure 1.3 which shows a switching node in an integrated environment, and Figure 1.4 that shows

the equivalent of the same section of the network in a non-integrated environment.

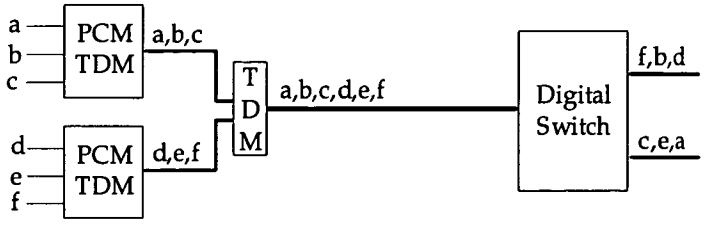


Figure 1.3: A digital switch in an Integrated Digital Network

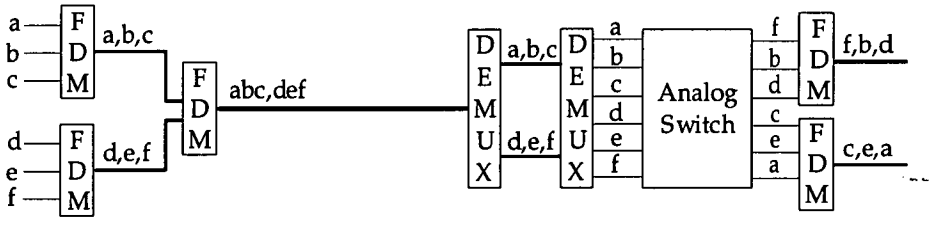


Figure 1.4: An analog switch in a non-integrated circuit switched network

The number of channels multiplexed and the levels of multiplexing in these figures are not realistic and have been chosen purely to give a simplified pictorial representation of these networks. As Figure 1.3 shows, because of TDM technology, the switching node along the path can employ a number of techniques to extract the time slots from the incoming link and switch the individual channels on the appropriate outgoing links. The equivalent process in the non-integrated analog network as shown in Figure 1.4 would require a lot of demodulation, demultiplexing, multiplexing and modulation, which would reduce the signal quality and increase the cost.

The cost advantage of Integrated Digital Networks is not only because of cheaper switching and transmission equipment. The other cost factor is the efficiency with which link capacities can be utilised. As the bandwidth of the transmission media increases (e.g. by employing optical fiber technology) the depth of multiplexing must be increased to use the available bandwidth on the link. This can be achieved more efficiently using TDM as compared to FDM.

Although Integrated Digital Networks were phased in primarily for achieving economical voice communication, they are also suitable for digital data communication. The integration of transmission and switching in a digital environment prepares the scene for integrating a wide range of data communication services with digitized voice, and leads to the next generation of telecommunication networks called the *Integrated Services Digital Networks* (ISDNs). Before describing an ISDN, the next two sections are devoted to local area networks (LANs) and metropolitan area networks (MANs). The reason for including these is that it is likely that interconnection of these networks will be an initial stage in the evolution of B-ISDNs.

1.2.4 Local Area Networks

A LAN is typically a collection of interconnected PCs, workstations, printers and data bases which are located within an organisation and in a relatively small geographical area (typically less than 10 km). In the industry a LAN may also involve interconnecting the above equipments with manufacturing systems. In a local area network many users share the same transmission medium. These days, the transmission speeds of such medium are at the edge of broadband speeds. Some examples are 10 Mbps Ethernets and 4 Mbps token rings. The more recent proposals for LANs call for even higher speeds, e.g. 16 Mbps token rings and 100 Mbps Fibre Distributed Data Interface [4]. One of the characteristics of LANs is their decentralised access control to the common medium. LANs are classified according to their topology, transmission medium and their medium access control (MAC) procedure. The most common types of LANs are token ring, token bus, and carrier sense multiple access with collision detection (CSMA/CD). A brief description of each of these networks follows.

- *Token Ring LAN:* is a closed loop (ring) which is made up of unidirectional point to point links interconnecting adjacent stations. It uses shielded twisted pair cable with transmission rates of 4 Mbps or 16 Mbps. The medium access control is based on the passing of a single token around the loop. When a station has some data to send it has to wait until it receives the token from its physical neighbour before it can transmit some packets.

One advantage of token protocols for MAC is that they prevent packet collision even under very high utilisations of the network. The drawback is that under lower utilisations the performance deteriorates because of the rotation time of the token.

- *Token Bus LAN:* consists of many stations which are passively coupled to the bus transmission medium. The transmission medium is a coax cable with transmission rates of 1 Mbps, 5 Mbps or 10 Mbps. A token passing protocol is used to control the access to the transmission medium. The token is passed from one station to its neighbour. A neighbour is defined by an address rather than by the physical location. The token passing protocol decides how many packets can be transmitted by a station while that station holds the token.

- *Carrier sense multiple access with collision detection (CSMA/CD):* uses a bus system with a transmission rate of 10 Mbps. It is based on the *Ethernet* LAN developed by Xerox. When a station has some packets to transmit, it listens for the carrier. If the channel is idle, the station will send the packets. If the channel is not idle, the station will wait for the channel to become idle before it transmits the packets. In this network it is possible for two stations to send packets simultaneously. This will result in collision. The transmitting stations will detect the collision and will stop transmission. Each of the stations will then wait for a random duration before trying to send packets again. This algorithm works very well for low network loading, but the performance of this network deteriorates with higher channel utilisations.

Local area networking is one of the areas where the high speed of broadband networks is needed. In many situations customers have LANs in different geographical locations and wish to interconnect their LANs. This interconnection may be achieved using existing networks such as circuit switched or packet switched data networks, or even ISDN. However, because the transmission rates of LANs are up to 16 Mbps, interconnecting them via slower networks will create a performance bottleneck in wide area networking. The high bandwidth local area networks require long distance broadband links to interconnect them [5]. The

high speed required for the interconnection of LANs may be achieved using the existing MANs or future B-ISDN. An example of interconnection of LANs via a MAN is shown in Figure 1.5.

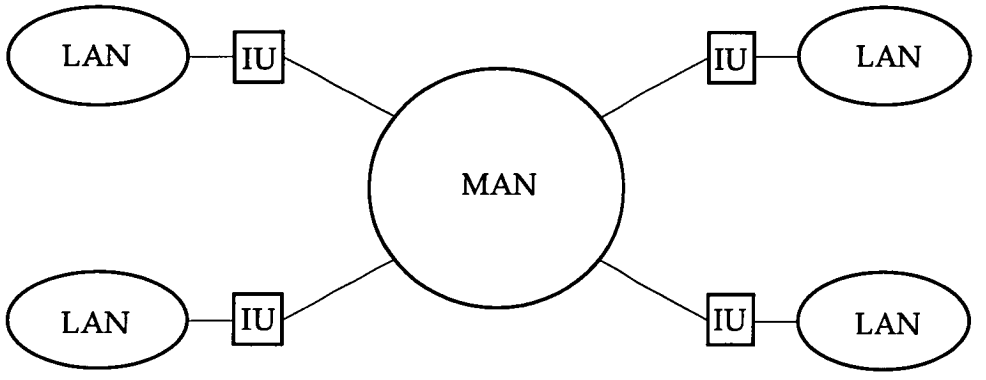


Figure 1.5: Interconnection of LANs via MAN (IU: Interworking Unit)

1.2.5 Metropolitan Area Networks

Metropolitan area networks (MANs) may be considered as an evolution of LANs. They provide fast data communication services in areas beyond local area networks. As mentioned earlier, MANs are the best present solution for interconnecting LANs. Some features of MANs are [6][7]: sharing a common transmission medium, geographical coverage of more than 50 km, distributed access control, high speed transmission ($\geq$ 100 Mbps), and provision for isochronous traffic (e.g. voice, video) as well as packet switched traffic.

MANs can be divided into two categories [8]: private, where a MAN is used by a single customer, and public, where many customers use the same MAN. In a public MAN the network operator must resolve such issues as billing, resource management and security. These issues are less complicated in a private MAN. Some alternatives for implementing MANs are fibre distributed data interface (FDDI) [9][10], fibre distributed data interface II (FDDI-II) [11][12], and distributed queue dual bus (DQDB) [13]. Details of these alternatives can be found in the citations and are omitted here.

In the same way that interconnection of LANs resulted in the evolution of metropolitan area networks, the interconnection of MANs (see Figure 1.6) will bring about the introduction of B-ISDN. The interconnection of MANs through B-ISDN will result in wider area access, improved performance and flexibility.

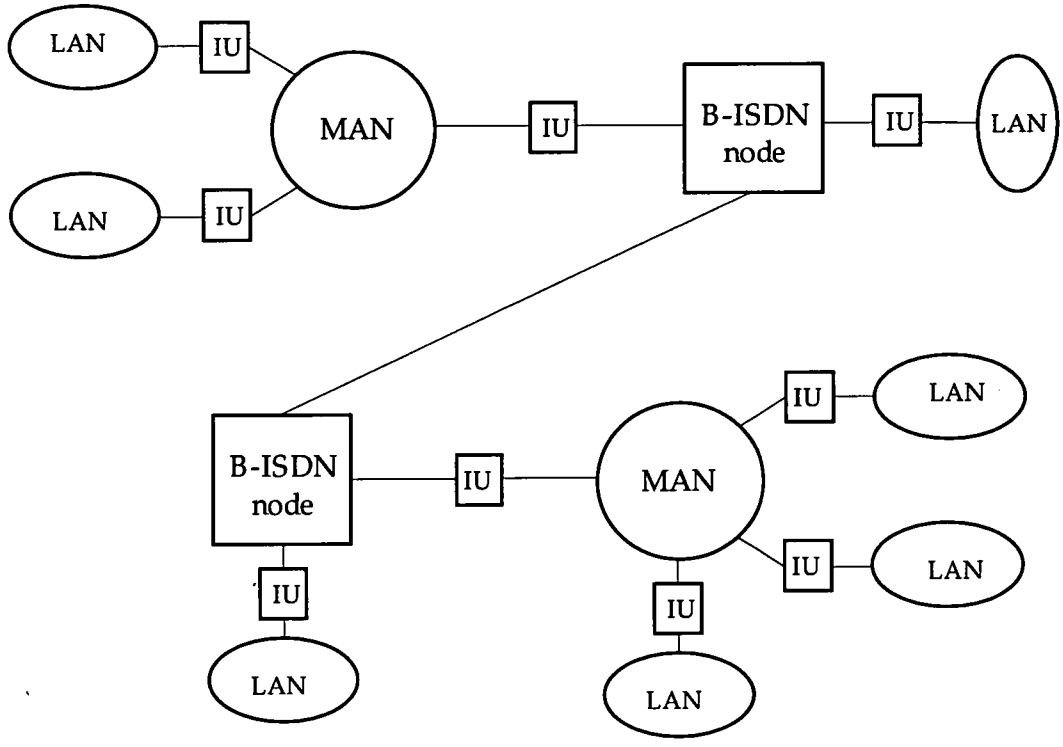


Figure 1.6: Interconnection of LANs and MANs via B-ISDN (IU: Interworking Unit)

1.2.6 Integrated Services Digital Networks

The term *Integrated* in Integrated Services Digital Networks refers to the integration of voice and data services on a single transport system. This will result in cost savings to users because they do not have to buy these services individually. Furthermore, the user can get access to a wide range of services via a single access line and a single standardized interface. This will particularly benefit the major corporate users that have experienced a rise in the cost of their separate voice and data networks [5]. Such customers are beginning to realise that an advanced,

integrated corporate network not only lowers their telecommunications costs but also provides them with additional functions that can give them a competitive edge in the market.

Let us now quote the principles on which ISDN is based from CCITT [14]:

" **1 Principles of ISDN**

1.1 *The main feature of the ISDN concept is the support of a wide range of voice and non-voice applications in the same network. A key element of service integration for an ISDN is the provision of a range of services (see Part II of the I-series of Recommendations) using a limited set of connection types and multipurpose user-network interface arrangements (see parts III and IV of the I-series of Recommendations).*

1.2 *ISDNs support a variety of applications including both switched and non-switched connections. Switched connections in an ISDN include both circuit-switched and packet-switched connections and their concatenations.*

1.3 *As far as practicable, new services introduced into an ISDN should be arranged to be compatible with 64 kbit/s switched digital connections.*

1.4 *An ISDN will contain intelligence for the purpose of providing service features, maintenance and network management functions. The intelligence may not be sufficient for some new services and may have to be supplemented by either additional intelligence within the network, or possibly compatible intelligence in the user terminals.*

1.5 *A layered protocol structure should be used for the specification of the access to an ISDN. Access from a user to ISDN resources may vary depending upon the service required and upon the status of implementation of national ISDNs.*

1.6 *It is recognized that ISDNs may be implemented in a variety of configurations according to specific national situations."*

The transmission structure of any ISDN link is constructed from three types of channels: B channel, D channel and H channel. These are described below.

- *B channel* has a capacity of 64 Kbps and can carry a wide variety of information streams except the ISDN signalling information. It is commonly used for PCM digitized voice and digital data.
- *D channel* has a capacity of 16 Kbps for basic access rate and 64 Kbps for primary access rate. The basic access rate and the primary access rate are defined later in this section. A D channel is primarily intended for carrying the common channel signalling information of the circuit switched calls in ISDN, but it can also be used for packet switched data and low speed telemetry.
- *H channels* are intended for higher bit rate user applications.

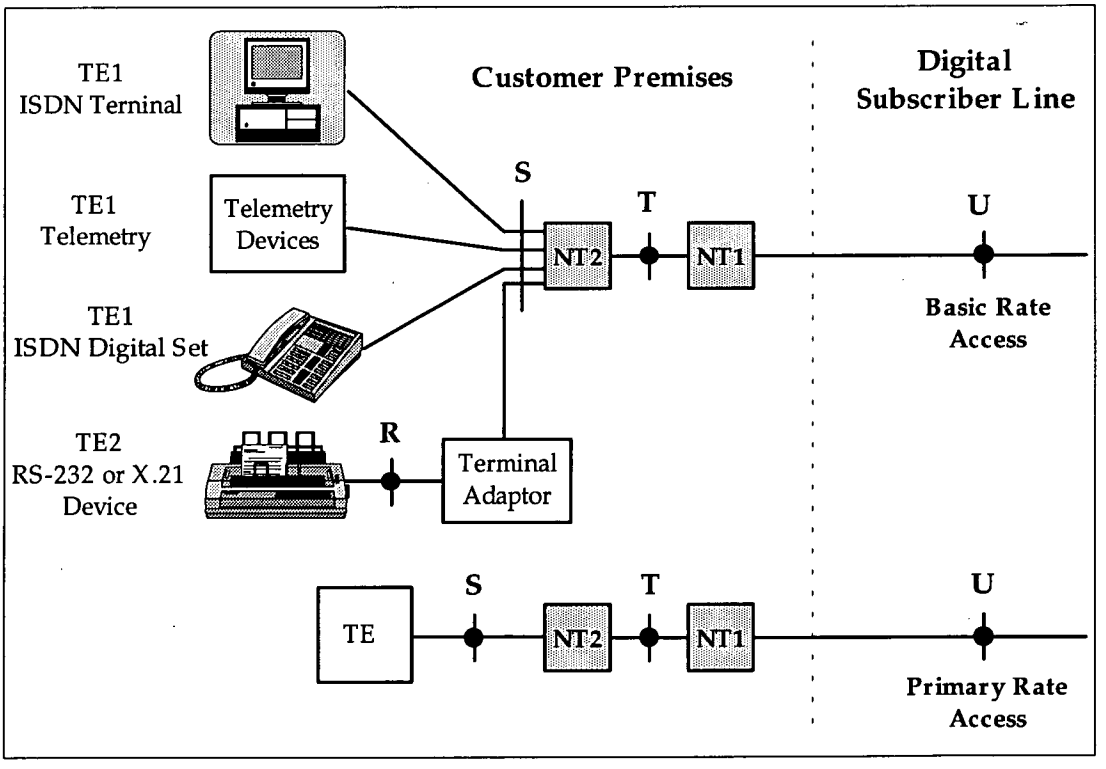


Figure 1.7: ISDN user-network reference model

Two well defined combinations of these channels, offered to users as a package, are the basic rate access and the primary rate access. The basic rate access con-

sists of two 64 Kbps B channels and one 16 Kbps D channel (total 144 Kbps). With overheads, the total bit rate of a basic access link is 192 Kbps. The basic rate access is intended for users with low bit rate demands such as residential customers. The primary rate in ISDN is intended for users with high bit rate requirements. The standard rate for primary access is 2.048 Mbps in Europe consisting of 30 B and 1 D channels. The primary access in the US, Canada and Japan is 1.5442 Mbps, organaised as 23 B and 1 D channels.

Figure 1.7 shows an ISDN user-network reference model showing reference points and functional groupings. ISDN functional groups are defined by sets of functions that may be needed in an ISDN access arrangement. NT1 (network termination type 1) includes necessary functions for physical and electromagnetic termination of the ISDN at the customer's premises. It isolates the customer premises equipments from the technology of the digital subscriber loop. NT2 (network termination type 2) is an intelligent interface between NT1 and TE (terminal equipment). It performs such functions as switching, concentration, multiplexing and protocol handling. A local area network is an example of a NT2.

At the customer's premises there may be three types of TE. These are TE1 (terminal equipment type 1), TE2 (terminal equipment type 2) and TA (terminal adaptor). TE1 includes ISDN terminals such as ISDN digital telephone set, ISDN PC terminal, ISDN telemetry devices, etc. TE1 connects directly to the S (system) reference point which is a 4-wire interface between a TE1 and a NT2. TE2 comprises of non-ISDN terminals, e.g. VT100, keysets, and RS232. It needs a TA to connect to the S interface. A TA's functions are rate adaptation of the lower speed data into 64 Kbps and bridging the existing non-ISDN products to ISDN. The R (rate) reference point provides an interface between adaptor equipment and non-ISDN user equipment. The T (terminal) reference point separates user's equipment from network provider's equipment at customer's premises. The U (user) reference point is a 2-wire interface for basic rate, or, a 4-wire interface for primary rate, between NT1 and an ISDN exchange.

Having quoted the basic principles of ISDN from CCITT documents it must be

pointed out that ISDN is not a fully integrated network. ISDN is only suitable for a limited range of services. For example, it can only offer services that require bit rates less than 2 Mbps. This excludes many video services from transmission over an ISDN. Even for narrowband services (less than 2 Mbps), although ISDN appears to be integrated as far as user access is concerned, it is not really integrated from network operator's point of view. ISDN still uses two different bearer services: packet switched and circuit switched, resulting in two overlay networks [15].

1.2.7 Broadband Integrated Services Digital Networks

The structure and details of ISDN have almost been completed with the publication of the *Blue Book* in 1988 by CCITT. Although further work is being carried out in refinement of the recommendations for ISDN, it is no longer the centre of attention amongst the researchers in telecommunications. Since 1988 CCITT has focused its attention on a far more complicated and revolutionary network concept known as the Broadband Integrated Services Digital Network (B-ISDN). One of the distinctions between ISDN and B-ISDN is the services offered by them. CCITT defines a broadband service as [16] *'A service or system requiring transmission channels capable of supporting rates greater than the primary rate.'* In recommendation I.121 [17], CCITT defines the principles of B-ISDN as following:

> " **2 Principles of B-ISDN**
>
> 2.1 *Asynchronous Transfer Mode (ATM) is the transfer mode for implementing B-ISDN and is independent of the means of transport at the physical layer.*
>
> 2.2 *B-ISDN supports switched, semi-permanent and permanent point-to-point and point-to-multipoint connections, and provides on demand reserved and permanent services. Connections in B-ISDN support both circuit mode and packet mode services of a mono- and/or multimedia type and of a connectionless or connection-oriented nature and in a bidirectional or unidirectional configuration.*

2.3 *The B-ISDN architecture is detailed in functional terms and is, therefore, technology and implementation independent.*

2.4 *A B-ISDN will contain intelligent capabilities for the purpose of providing advanced service characteristics, supporting powerful operation and maintenance tools, network control and management. Further inclusion of additional intelligent features has to be considered in an overall context and may be allocated to different network/terminal elements.*

2.5 *Since the B-ISDN is based on overall ISDN concepts, the ISDN access reference configuration is also the basis for the B-ISDN access reference configuration.*

2.6 *A layered structure approach, as used in established ISDN protocols, is also appropriate for similar studies in B-ISDN. This approach should also be used for studies on other overall aspects of B-ISDN including information transfer, control, intelligence and management.*

2.7 *Any extension of network capabilities or change in network performance parameters will not degrade the Quality of Service of existing services.*

2.8 *The evolution to B-ISDN should ensure the continued support of existing interfaces and services.*

2.9 *New network capabilities will be incorporated into B-ISDN in evolutionary steps to meet new user requirements and accommodate advances in network developments and progress in technology.*

2.10 *It is recognised that B-ISDN may be implemented in a variety of ways according to specific national situations."*

Although the long term goal of B-ISDN would be to provide all types of services, it is likely that initially it will only be used for those services that cannot be offered (cost effectively) by the existing networks. B-ISDN would have to coexist with the networks already in the place such as the Public Switched Telephone Networks (PSTNs), public data networks, narrowband ISDN, etc. until these technologies are phased out by time.

1.3 B-ISDN Services

B-ISDN is expected to offer a wide range of services from very low bit rate data services such as telemetry to very high bit rate video services such as High Definition Television (HDTV). It is not easy to predict all possible services in a B-ISDN because many services supported by B-ISDN will be value added services. Such services may be offered by companies other than the network providers. An independent company may think of an economically viable new service and use B-ISDN to offer it.

Although an exact picture of B-ISDN services is not at hand, it is possible to identify the categories of the possible services. CCITT has identified two main service categories in a B-ISDN: interactive services and distribution services. These are shown in Figure 1.8 [18].

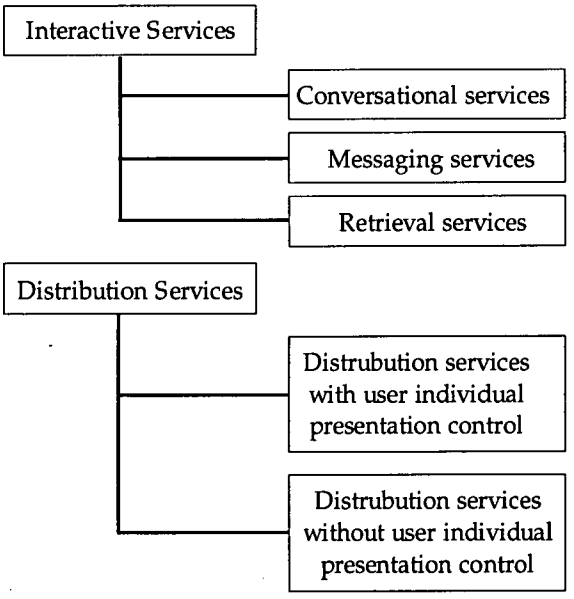


Figure 1.8: Classification of broadband services

1.3.1 Interactive Services

Interactive services, as the name implies, refer to those services in which there is a bidirectional transfer of data between two points. Signalling and control data are

not included in this definition. Interactive services are point to point services, i.e. between a service provider and a subscriber or between two subscribers. CCITT defines three classes of service that come under the interactive services category. These are conversational services, messaging services and retrieval services.

- *Conversational services* are services where there is a real time (i.e. no store and forward) bidirectional end-to-end information transfer between two users or between a user and a host. Some examples of conversational services are broadband video telephony, broadband video conferencing, video surveillance, high volume file transfer and high speed telefax.

- *Messaging services* are user to user communication services via storage units with store-and-forward, mailbox and/or message handling functions. Video mail and document mail are two examples of messaging services.

- *Retrieval services* are services where information is stored in information centres and users can selectively retrieve those information, i.e. information will be sent to a user only on demand. The user can control the time when an information sequence starts. Some examples of retrieval services are broadband videotex, video retrieval services, high resolution image retrieval services and document retrieval services.

1.3.2 Distribution Services

Distribution services provide point to multipoint communication and are generally from a service provider to a large number of users. CCITT defines two classes for distribution services. One class is distribution services without user presentation control, and the other class is distribution services with user presentation control. A brief description of each of these classes is given below.

- *Distribution services without user individual presentation control* include broadcast services. The service provider provides a continuous flow of information that is distributed to many authorised users. The user can get access to this flow of information any time, but he can only access the information that is broadcasted from the time of his connection onward. The user cannot control the time and order of the presentation. An example is

broadcast television which at the moment is only accessible to users through microwave or cable, but in a B-ISDN environment can be integrated with the rest of the services.

- *Distribution services with user individual presentation control* are also services that distribute information from a central source to many users. However, the information is provided as a sequence of information entities with cyclical repetitions. Therefore the user can get access to individual information entities and can also control the start and the order of the presentation. An example is cabletext which is an enhanced, broadband version of teletext.

1.4 B-ISDN Protocol Reference Model

In recommendations I.320 [19] and I.321 [20], CCITT describes a protocol reference model (PRM) for B-ISDN which reflects the principles of layered communication of the reference model of open system interconnection (OSI) in recommendation X.200 [21]. Most principles of OSI including protocol layering, layer service definition, service primitives and modularity seem appropriate to the B-ISDN environment. However, the OSI principle of layer independence has not been fully applied to the B-ISDN protocol reference model.

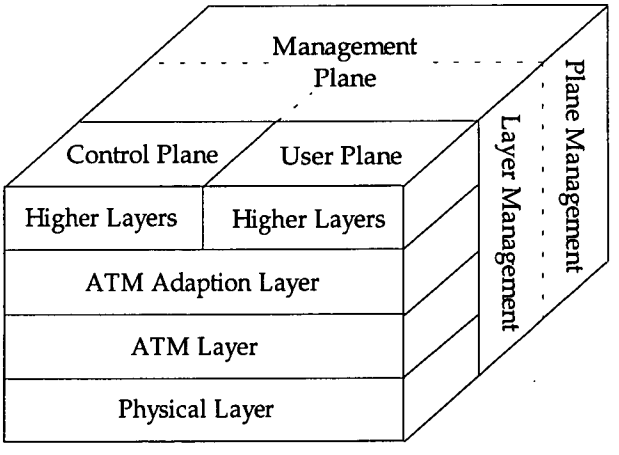


Figure 1.9: B-ISDN protocol reference model

1.4.1 Planes of B-ISDN Protocol Reference Model

Figure 1.9 [20] shows the protocol reference model for B-ISDN. There are three planes in the PRM of B-ISDN:

- The *User plane* has a layered structure and provides for user information transfer, along with associated controls such as flow control and error control.
- The *Control plane* with its layered structure provides the call control and connection control functions.
- The *Management plane* includes layer management and plane management functions:
  - *Plane management* performs management functions to a system as a whole and provides coordination between all the planes.
  - *Layer management* performs management functions relating to resources and parameters residing in its protocol entities.

1.4.2 Layers of B-ISDN Protocol Reference Model

In this section the functions of various layers and the primitives exchanged between them are discussed. Figure 1.10 [20] shows the functions of B-ISDN in relation to protocol reference model. A brief description of each layer follows.

Physical Layer

The physical layer consists of two sublayers: the physical medium (PM) sublayer and the transmission convergence (TC) sublayer.

- *Physical medium sublayer* must perform all necessary functions such as bit transfer, bit alignment, line coding and electrical to optical transformation to provide bit transmission capability. Specification of the physical medium sublayer will depend on the medium used for transmission.
- *Transmission convergence sublayer* performs all necessary functions to convert the flow of cells to flow of data units that can be transmitted over

<table>
<tr><td rowspan="5">Layer Management</td><td>Higher Layer Functions</td><td colspan="2">Higher Layers</td></tr>
<tr><td>Convergence</td><td>CS</td><td rowspan="2">AAL</td></tr>
<tr><td>Segmentation & reassembly</td><td>SAR</td></tr>
<tr><td>Generic flow control
Cell header generation/extraction
Cell VPI/VCI translation
Cell Multiplex & demultiplex</td><td colspan="2">ATM</td></tr>
<tr><td>Cell rate decoupling
HEC header sequence generation/verification
Cell delineation
Transmission frame adaption
Transmission frame generation/recovery</td><td>TC</td><td rowspan="2">Physical Layer</td></tr>
<tr><td>Bit timing
Physical medium</td><td>PM</td></tr>
</table>

Figure 1.10: Functions of B-ISDN protocol reference model layers

the physical medium. These functions are transmission frame generation and recovery, transmission frame adaptation, cell delineation, header error control sequence generation and cell header verification, and cell rate decoupling. These functions are explained below:

- *Transmission frame generation and recovery* as the name explains, performs the functions required for generation and recovery of transmission frame.

- *Transmission frame adaptation* must provide all necessary conversions between cell flow and the transmission frame. In the transmit direction it must structure the flow of cells according to the payload structure of the transmission frame. In the receive direction it must extract the cell flow from the payload of the transmission frame.

- *Cell delineation* maintains the cell boundaries to enable the receiving end to identify the cells after descrambling the ATM cell stream. The self delineating mechanism for this purpose has been defined in recommendation I.432 [22].

- *HEC sequence generation and cell header verification:* Each ATM cell has a Header Error Control (HEC) field. This function is responsible for calculating the HEC field in the transmit direction. In the receive direction, cell headers are checked for errors and those errors that can be corrected are corrected. Those cells with non-correctable corrupted cell headers are discarded.
- *Cell rate decoupling:* Because of asynchronous nature of the ATM cells, there is a need to adapt the rate of the valid ATM cells to the rate of the payload of the transmission frame. This is achieved by insertion and suppression of idle cells.

ATM Layer

The ATM layer is independent of the physical medium used to transport the ATM cells and therefore independent of the physical layer. The ATM layer must perform four main functions as follows.

- *Cell multiplexing and demultiplexing:* For the purpose of transmission, the cells from individual Virtual Paths (VPs) and Virtual Channels (VCs) are combined into a non-continuous composite cell flow. At the appropriate point in the receiving end, cells are extracted from the composite cell flow and are directed to appropriate Virtual Paths or Virtual Channels.
- *Virtual Path Identifier (VPI) and Virtual Channel Identifier (VCI) translation:* At ATM switching and/or cross connect nodes, the values of the VPI and the VCI of the incoming ATM cells must be translated into new values of VPI and VCI on the outgoing links.
- *Cell header generation/extraction:* In the transmit direction this function must generate appropriate ATM cell headers (except the HEC sequence) for the cell information fields received from ATM Adaptation Layer (AAL). This function may also include the translation from a Service Access Point (SAP) identifier to a logical connection number (VPI and VCI).
- *Generic Flow Control:* This function is responsible for generating a Generic Flow Control (GFC) field for the ATM header.

Further aspects of the ATM layer will be discussed in Section 1.5.

ATM Adaptation Layer

There are some applications that cannot use the services of the ATM layer directly. Therefore, the ATM Adaptation Layer (AAL) is necessary to enhance the services of the ATM layer for specific applications. The AAL lies between the ATM layer and higher layers. It must adapt the services provided by the ATM layer to the requirements of the higher layers. Functions of the ATM Adaptation Layer can be divided into two sublayers: the *segmentation and reassembly sublayer* (SAR), and the *convergence sublayer* (CS). In the transmitting side the SAR must segment the higher layer Protocol Data Units (PDUs) into the size of the ATM cell information field (48 octets). At the receiving side, the SAR reassembles the information field of the ATM cells into the PDUs of higher layer. The convergence sublayer is service specific and defines the services that AAL must provide for the higher layers. For example, for high quality audio and video services, CS may provide forward error correction to protect against bit errors. It may perform other functions for other services.

1.5 B-ISDN and Asynchronous Transfer Mode

Asynchronous Transfer Mode (ATM) is the transfer mode solution recommended by CCITT for implementing B-ISDN [23]. It results from the merging of two well-known concepts: packet switching and time division multiplexing [2] (these topics have been discussed earlier in this chapter). It is a specific packet transfer mode that uses asynchronous time division multiplexing. ATM is a connection oriented technique. Connection identifiers are assigned to each link of the connection and are maintained for the duration of the call. ATM can support all services, including connectionless services. The ATM layer maintains the cell sequence integrity on a virtual channel connection and the higher layers (e.g. AAL) provide additional functions necessary for supporting different services. The following section explains the ATM transport network and includes formal definitions of Virtual Channel and Virtual Path which are used frequently in this chapter.

1.5.1 The ATM Transport Network

Figure 1.11 [24] shows the transport hierarchy of an ATM network. It consists of two layers: the ATM Layer and the Physical Layer. The ATM layer of the transport network has two levels, the Virtual Channel level and the Virtual Path level. A Virtual Channel (VC) is a logical connection in an ATM network and provides user to user communication capability for the transport of the ATM cells. It also provides network signalling capability between user and network, and network management and routing capability between network and network.

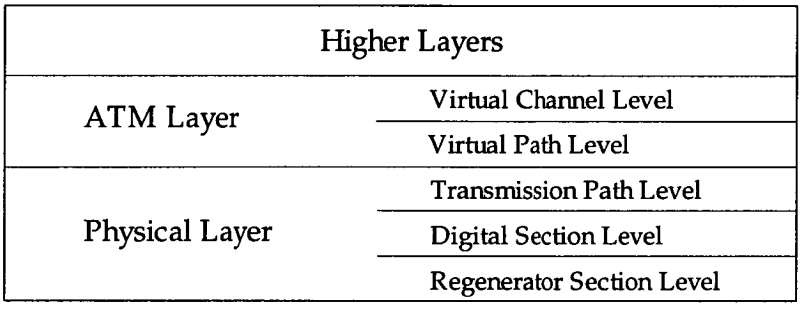


Figure 1.11: ATM Transport Hierarchy

A Virtual Channel is the smallest unit in the switching nodes of B-ISDN. A Virtual Channel Identifier (VCI) is assigned to the header of all cells belonging to the same virtual channel on a link. The value of the VCI is changed every time that a VC is switched. The concatenation of all VC links that provide end to end communication capability is called a Virtual Channel Connection (VCC). A VCC can provide user-user, user-network, or network-network information transfer capability.

A Virtual Path (VP) is a bundle of Virtual Channels that have the same end points, i.e. all the channels in a Virtual Path have the same Virtual Channel Identifier. A specific value of Virtual Path Identifier (VPI) is used to group all the VC links that share the same Virtual Path Connection (VPC). A new value of VPI is assigned every time that a VP is switched. Virtual Paths are considered a substantial component of a resource management control hierarchy for the B-

ISDN [25] (The topic of traffic control and resource management will be discussed in section 1.6). The relationship between VC, VP and transmission path is shown in Figure 1.12 [24].

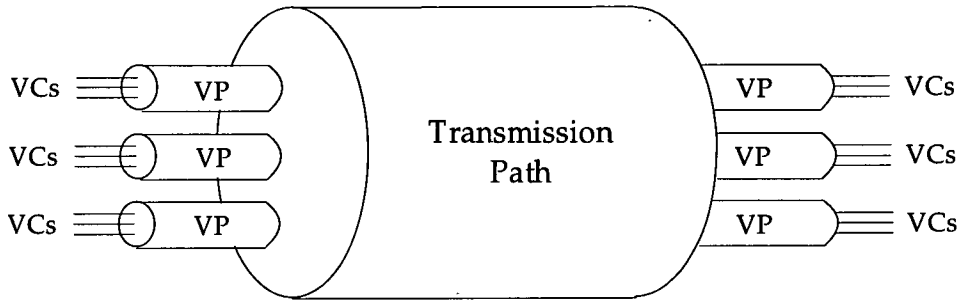


Figure 1.12: The relationship between VC, VP and Transmission Path

The physical layer of the transport network has three levels: transmission path level, digital section level and regenerator section level. Details of the physical layer have been omitted here. A description of this layer is given by CCITT in recommendation I.311 [24].

1.5.2 ATM Cell Structure

In a packet switched network, two parameters can be varied in relation to the mode of transport. These are packet length, within some limits, and the time between packets. Traditional packet switching networks vary both the size of the packets and the time between them. ATM on the other hand only varies the time between packets (cells) and keeps the size of the packets (cells) fixed.

In Asynchronous Transfer Mode, information is divided into 48 octet segments and a 5 octet header is assigned to each segment to build a fixed size cell of 53 octets as shown in Figure 1.13. The header can have two formats depending on the interface: the user-network interface format and the network-node interface format. These formats are shown in Figures 1.14 and 1.15 [26]. The rest of this section is devoted to the description of fields of the cell header.

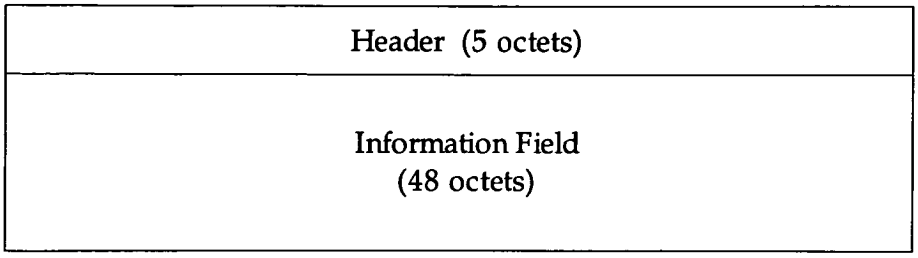


Figure 1.13: ATM cell structure

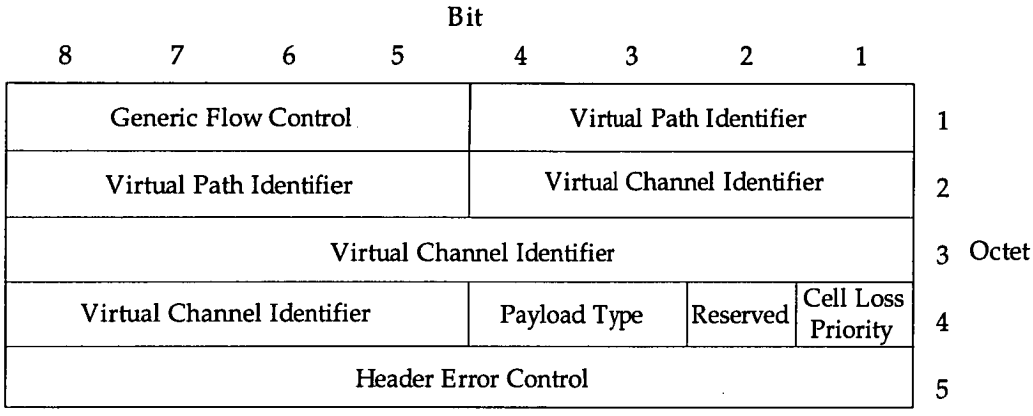


Figure 1.14: ATM cell header at user-network interface (UNI)

- *Generic Flow Control (GFC) field* only applies to the user-network interface and consists of 4 bits. It will be used for end-to-end flow control. The functions of GFC have not yet been standardised and until then a default value of 0000 will be used for this field. This field may be used to assist the customer in defining multiple priority levels for controlling the quality of service for different traffic types.

- *VPI field* is used for identification of different Virtual Path links that have been multiplexed at the ATM layer into the same Physical Layer connection at a given interface and a given direction. At the user-network interface this field consists of 8 bits and is used for routing. At the network-node interface it consists of 12 bits and provides enhanced routing capabilities. The exact number of bits of the VPI field used for routing is established by negotiation between the user and the network, using rules defined in recommendation I.361 [26]. All bits of the VPI field are set to zero in an unassigned cell .

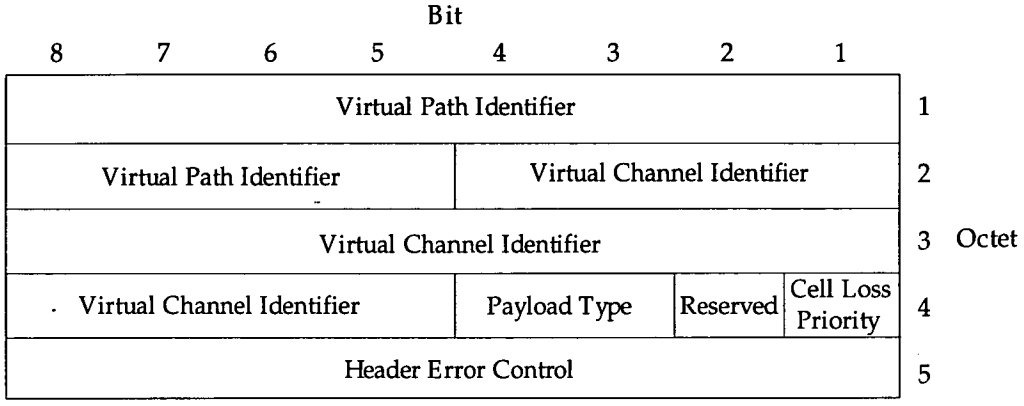


Figure 1.15: ATM cell header at network-node interface (NNI)

- *VCI field* is used to distinguish the different Virtual Channel (VC) links in a Virtual Path Connection (VPC). It consists of 16 bits for both user-network and network-node interfaces.

- *Payload Type (PT) field* consists of 2 bits. The default value of this field is 00 for cells that carry user information. Other values of PT for user information and for network information are for further study.

- *Reserved field* is second bit of octet 4 of the header. The exact use of this field has not yet been specified, but, it is expected that this field will be used to enhance the existing cell header functions or for standardised functions not yet specified. The default value for this field is 0.

- *Cell Loss Priority (CLP) field* is 1 bit and explicitly indicates the priority of the cell. Cells that have a CLP value of 0 are of higher priority. If the CLP value is 1, the cell may be subject to discard depending on the network condition. The negotiated Quality Of Service (QOS) will not be violated by the discard of these cells, should it become necessary. The assignment of CLP value may be made by the user or by the service provider. In Constant Bit Rate (CBR) services all the cells have high priority. For Variable Bit Rate (VBR) services however some cells may be labelled low priority and other cells that are essential for maintaining the guaranteed QOS will be labelled high priority.

- *Header Error Control (HEC) field* is the last octet of the header and covers the entire cell header. It can detect and correct single-bit errors, and can detect (only) multiple-bit errors. A description of the mechanism of HEC is given in recommendation I.432 [22].

1.6 Traffic Control and Resource Management

A B-ISDN has many attractive features including flexibility to support a spectrum of service types and mixes on demand, efficiency gained by statistically multiplexing bursty traffic generated from several sources and service types, and simplicity by avoiding the need for multiple operations, administration and maintenance systems [27]. None of the preceding networks have been capable of providing all of these features. There is however a price to pay for these attractive features of B-ISDN and that is the additional traffic control and resource management complexity.

There are many factors that render the traffic control strategies designed for traditional packet switching networks unsuitable for B-ISDN. The traditional packet switching networks (e.g. X.25) are designed for homogeneous traffic environments where it is acceptable for all packets to be treated equally while competing for network resources. Error and flow control in such networks are achieved by link-by-link window flow control, but there are no controls geared explicitly towards meeting grade of service objectives [28]. B-ISDN provides the ability to support a wide range of services (new and old) but it also introduces new challenges concerning traffic control and resource management. Statistical multiplexing in B-ISDN leads to efficient use of network resources but it requires new methods of traffic control and bandwidth management [29][30].

In the traditional packet switched networks, the performance bottleneck has mainly been the channel transmission speed. Due to the high speed of broadband networks, channel transmission speed is no longer an important issue of network performance. As the transmission rate increases the performance bottlenecks are shifted to processing speed at the network switching nodes and the propagation

delay of the channel. Due to large bandwidth-propagation delay product, large number of cells can be in transit between two nodes. The result of this is that traditional reactive flow control schemes will be less effective in detecting and reacting to congestion [31]. Also, due to the increased ratio of processing time to propagation delay it is difficult to implement hop-by-hop control schemes.

	Voice	*Bulk Data*	*Interactive Data*	*Video*
Bit Rate	64Kbps	8Mbps	5Mbps	0.064 - 200 Mbps
Pattern	Stream	Stream	Bursty	Stream
Holding Time	Minutes	Minutes	Minutes	Minutes
Bit Error Rate	10^{-3}	10^{-12}	10^{-10}	10^{-6}
Delay Tolerance	Milliseconds	Seconds	Seconds	Milliseconds
Sensitivity to variable delay	Limited	Insensitive	Limited	Severe

Table 1.1: Typical Traffic Characteristics

Let us consider some typical characteristics of a few services in B-ISDN as shown in Table 1.1. While it is clear that with heterogeneous traffic mixtures there may be conflicting quality of service requirements, the aim of the network control should be to give each type of traffic an acceptable grade of service, judged by the criteria appropriate for that traffic type. Unstated among the performance criteria shown in Table 1.1 is network operator's need for control schemes that have low overheads, are simple to implement, allow high utilizations of network resources, and can deliver the required levels of performance without too much retuning as the network load and traffic mixes vary.

Let us quote CCITT's definition of traffic control and resource management: *'The primary role of traffic control and resource management parameters and procedures is to protect the network in order to achieve network performance objectives. An additional role is to optimise the use of network resources'* [32].

The various control approaches for achieving these objectives can be divided into two main categories [27]: reactive and preventive. These concepts are discussed here.

- *Reactive Control* regulates the traffic entering the network, based on the current traffic levels in the network. This means that the congested node immediately instructs the source nodes to limit their traffic flow. In a high speed network there is a major problem with this type of traffic control. It is very difficult to detect the onset of congestion well in advance to react effectively. By the time that the message of the congested node reaches the source nodes and is processed, the high capacity links would have greatly suffered. This type of control is therefore more suitable for private, localised networks carrying homogeneous traffic, where all end terminals or systems can be throttled back in a similar way, and where the network users can be depended upon to protect the integrity of the network [28].

- *Preventive Control* ensures that the network traffic will very rarely reach the level at which congestion results. In other words congestion is considered a rare event. This approach is more suitable for integrated transport. Preventive control may be implemented in two ways: by over-engineering the network, in which case there should be no need for control, or by controlling the traffic flow admitted into the network at the access nodes. This latter method obviously uses the network resources more efficiently.

It is likely that initial dimensioning and call acceptance control of B-ISDN will be based on peak rate allocation [33], meaning that the network will be over-dimensioned so that it can cope with simultaneous peak rate transmission of all connections. The latter phases of B-ISDN will aim at more efficient use of network resources. At that stage the traffic control schemes are unlikely to be as black and white as reactive control *'or'* preventive control as outlined earlier, rather, they will be a combination of traffic control functions in both categories, perhaps with more emphasis on preventive measures.

An ATM network can provide several levels of traffic control capabilities as follows [24]:

- Connection admission control
- Usage parameter control
- Priority control
- Congestion control

Each of these levels tackles a particular set of problems in relation to traffic control and it is important that a high level of consistency exists between these levels. The following subsections give a description for each of these traffic control levels and provide a literature survey as appropriate.

1.6.1 Connection Admission Control

According to CCITT, *'Connection admission control is defined as the set of actions taken by the network at the call set up phase (or during call renegotiation phase) in order to establish whether a (virtual channel/virtual path) connection can be accepted or rejected'* [24].

When the network receives a connection request, the connection admission control scheme must decide if sufficient network resources are available to satisfy the performance requirements (e.g. acceptable cell transmission delay and cell loss probability) of the new connection while maintaining the agreed quality of service of the connections already established. The signalling messages sent by the user at the time of the connection request should contain information about the source traffic characteristics and about the required quality of service class. The main issues in connection admission control are:

- The traffic descriptors that can accurately describe the traffic that will be generated by the new service.
- The relationship between traffic descriptors and network performance.
- The decision criteria for connection admission control.

A discussion of these issues follows.

Traffic Descriptors

At the time of a connection request, the user must provide the network a set of parameters that describe the traffic characteristics of the requested connection. Ideally, this set should contain the smallest number of parameters that can be used to predict accurately the ability of the network to maintain a certain level of performance. Network performance should be insensitive to other source characteristics [34]. Some parameters that can be used in characterising the source traffic are:

- average bit rate
- peak bit rate
- burstiness
- peak duration.

The precise quantitative definition of burstiness is still a pending issue in CCITT but it can generally be defined as a measure of how densely or sparsely cell arrivals occur. There are several definitions proposed for burstiness, e.g. the ratio of peak bit rate to mean bit rate [35, 36, 27], burst factor defined as (peak bit rate - mean bit rate) $\times$ average burst length [37, 38], peakedness defined as the variance-to-mean ratio of the number of busy servers in a fictitious infinite server group [39], cell jitter ratio defined as the variance-to-mean ratio of the cell interarrival time [40], and the squared coefficient of variation of the interarrival times [41].

Some traffic descriptors given above are correlated. For example, the mean bit rate and the peak bit rate are correlated with burstiness. There are other traffic descriptors that have been proposed, such as the effective bit rate defined as a fraction of the peak bit rate [42].

The Decision Criteria for Connection Admission

The most commonly used performance parameters used as decision criteria for connection admission are cell loss probabilities and cell transmission delays. In

the past many researchers have used the long term time-averaged values of these parameters [36, 42, 43] as the decision criteria. These long term time-averaged values are not sufficient performance measures in an ATM network because in an ATM network the volume of the traffic can change very rapidly. The ineffectiveness of using the long term time-averaged cell loss probability as the decision criterion for connection admission has been demonstrated in [44].

With the bursty nature of many ATM traffic sources, it is possible to get short term congestion where large number of cells are lost, to the extent that the service quality is not acceptable to the user although the cell loss probabilities averaged over the long term are still very small [44]. In [44] it is suggested that instantaneous cell loss probabilities be used as the decision criterion to study the short term behaviour of a network. These probabilities are approximated by steady state values. In the method proposed by Kamitake and Suda [44], when a connection request is made it will only be accepted by the network if the instantaneous cell loss probability is below a threshold value for a predetermined percentage of time.

The inadequacy of the long term time-averaged cell loss probabilities have also been shown in [45] where the short term behaviour of cell loss probability of voice cells are studies. It is reported that with realistic sets of parameters the cell loss rate changes slowly and remains at zero for most of the time. However at the onset of a congestion the cell loss probability becomes very large to the extent that the recipient can detect the distortion in the voice. Therefore, although the long term time-averaged cell loss probability is very low it does not reflect the unacceptable cell loss rate within a blocking period (i.e. the period of time during which the buffer is full).

The Effect of Traffic Descriptors on the Network Performance

This is an important issue in connection admission control. There are several examples in literature which investigate the effects of statistical multiplexing of several bursty sources in an ATM network to determine how the network performance varies as a function of different traffic parameters. Most of the

findings are intuitively obvious. Some of these findings are:

- The cell loss probability increases as the peak rate of each source is increased [37, 43].
- As the burst length is increased the cell loss probability and mean delay increase significantly [37, 43, 36].
- The cell loss probability increases as the offered load increases [36].
- When several homogeneous sources are multiplexed, with a constant offered load (defined as the number of sources $\times$ mean bit rate of each source), the cell loss probability decreases as the number of sources increases. The reason is that when the number of sources is increased (constant offered load) the mean bit rate must decrease. Therefore, either or both of the peak bit rate and the peak duration must decrease. This results in a reduction in cell loss probability [36, 37].
- When heterogeneous sources are multiplexed, the fluctuation in the cell loss is greater when high bit rate sources have been multiplexed. The reason is that when low bit rate sources are multiplexed, the large number of the sources will have a smoothing effect on the multiplexed traffic [44].

1.6.2 Usage Parameter Control

According to CCITT, *'Usage parameter control is defined as the set of actions taken by the network to monitor and control (user's) traffic in terms of traffic volume and cell routing validity. Its main purpose is to protect network resources from malicious as well as unintentional misbehaviour which can affect the QOS of other already established connections by detecting violations of negotiated parameters'* [24].

Connection admission control bases its decision on the traffic parameters supplied by the user at the connection set-up time. Obviously, users may intentionally or unintentionally change their traffic beyond what is negotiated at the call set up time. The complement to access control is usage parameter control. After a

connection has been established, the traffic flow from the connection should be monitored to ensure its conformity to the traffic descriptors supplied by the user. The monitoring algorithm must be implemented at appropriate points in the network. It may be performed, depending on the customer access configuration, on virtual circuits, on virtual paths, or on the total traffic volume on an access link within components like concentrators, local exchanges, and ATM cross-connects [46]. When there are two or more network operators, traffic monitoring functions may also be necessary at the boundaries of the subnetworks belonging to different operators.

The usage parameter control must somehow identify the voilating traffic and must also have a strategy for dealing with connections that violate the traffic parameters established for them at the connection set up time. There are a number of mechanisms that the usage parameter control may implement to achieve its objectives. Some mechanisms are: the leaky bucket algorithm [47], the jumping window mechanism [46], the triggered jumping window mechanism [46], the exponentially weighted moving average mechanism [46, 48], the moving window mechanism [46], the rectangular sliding window [48], and the triangular sliding window [48]. Some of these algorithms are explained below.

Leaky Bucket Algorithm

The Leaky Bucket (LB) algorithm enforces the average bit rate and burstiness of a source. This method is considered one of the most promising bandwidth enforcement strategies. The LB algorithm was first proposed by Turner [47] and since then it has been investigated and refined by several authors [49, 50, 51, 52, 35].

One method of implementing the LB is by means of tokens as conceptually shown in Figure 1.16. For a cell to enter the network it must obtain a token from the token pool. Tokens are generated to the token pool with a frequency corresponding to the mean bit rate of the source as established during call set up time. When a cell is generated from the source it must enter the queue. The queue is served on a FIFO basis by the tokens in the pool. For each cell that leaves the queue there must be a token in the pool. When a cell leaves the queue the number of

tokens in the pool is decremented by one. In the event of a full queue, any new incoming cell is discarded. In the case of an empty pool, cells must wait in the queue until more tokens are generated.

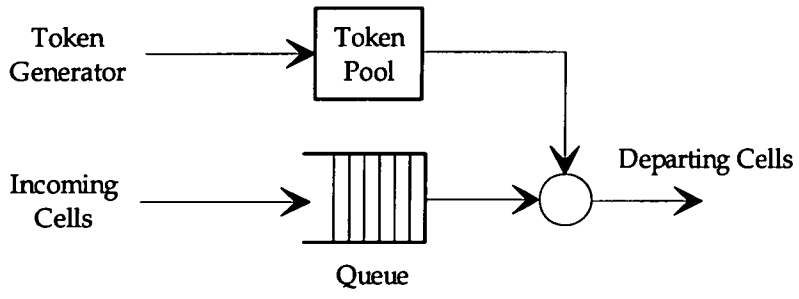


Figure 1.16: A token controlled leaky bucket method

The physical realisation of a LB corresponds to a counter that is incremented every time that the source generates a cell. This counter is decremented periodically with an appropriate rate. A threshold value, N, is defined for the counter. A cell arriving when the counter has reached the threshold value is dropped, or tagged as a violating cell [35, 50].

Jumping Window Algorithm

The Jumping Window algorithm (JW) [46] divides the time into fixed intervals called windows. A source can transmit a maximum of N cells during each window. The new window starts immediately after the end of the preceding window (hence the name jumping window) and the corresponding cell counter is reset to zero at the beginning of the new window. Similar to LB algorithm, counters are needed to keep track of the time and the number of cells. In order to calculate the probability of traffic violation, the arrival process must be used to derive the probability distribution of the counting process. Then the first N elements of the probability distribution of the counting process are necessary to calculate the probability of traffic violation.

Triggered Jumping Window Algorithm

The Triggered Jumping Window (TJW) algorithm is similar to JW, except that the windows are not consecutive but are triggered by the first cell arriving since the end of the last window.

Exponentially Weighted Moving Average Algorithm

The Exponentially Weighted Moving Average (EWMA) [46] also uses the concept of limiting the number of cells arriving during each window. The difference between this algorithm and the JW algorithm is that the maximum number of cells accepted in the i^{th} windows, N_i, is both a function of the mean threshold value N, and the number of cells accepted in the preceding intervals (X_i) according to rule [46]

$$N_i = \frac{N - \gamma S_{i-1}}{1 - \gamma} \quad 0 \leq \gamma < 1$$

with

$$S_{i-1} = (1 - \gamma)X_{i-1} + \gamma S_{i-2}$$

where S_0 is the initial value of EWMA measurement and γ is the flexibility factor to the burstiness of the traffic. This algorithm is obviously more complicated to implement compared to the earlier algorithms (LB and JW).

Moving Window Algorithm

The Moving Window (MW) algorithm is again similar to the JW algorithm where the maximum number of cells accepted during each window is limited to N. The difference is that now the window is steadily moving along the time axis. Each cell is remembered for exactly one window interval. When a cell is accepted, it's arrival time is stored and the counter is incremented. Exactly T time units later (T is the duration of a window), the counter is decremented by 1. The complexity of this algorithm is a function of N, the maximum number of cells accepted during each window, because it must keep the record of the arrival time of up to N cells in each window. Therefore this algorithm is relatively more expensive to implement for a realistic problem.

In [46] it is concluded that Out of the algorithms described above, LB and EWMA are the most promising mechanisms because the other mechanisms are not flexible enough in coping with the short term statistical fluctuations of the source traffic.

It must be noted that while the usage parameter control should monitor the traffic generated from a source for its conformity to the traffic descriptors negotiated during connection establishment, the control traffic characteristics may still be different from the negotiated values due to such factors as faulty network elements or faulty control devices. In such cases the management plane should take the necessary actions to overcome the fault.

1.6.3 Priority Control

As stated earlier, the ATM cell header has a one-bit cell loss priority (CLP) field which explicitly indicates the priority of the cell. It is possible to make use of this field for the cells belonging to the same virtual channel or virtual path if the information to be transmitted can be classified into more important and less important categories. In that case the more important parts of the information may be packed into high priority cells and the less important information can be packed into low priority cells. The connection admission control and the usage parameter control can then treat the two categories of cells separately.

The priority control may be implemented at the intermediate buffers by designing strategies that would decide how the high priority and low priority cells are buffered. Several papers have studied buffer priority mechanisms [53, 54, 55, 56, 57, 58, 59]. Kröner [53] outlines and compares three different buffering mechanisms. These mechanisms are:

- *Common buffer with pre-emption:* with this mechanism high priority and low priority cells share the same buffer. If a high priority cell encounters a full buffer and there are one or more low priority cells in the buffer, then one low priority cell will be pre-empted and lost to free up space for the high priority cell. For this buffering mechanism complex buffer management

would be necessary to ensure that cell sequence integrity is preserved.

- *Partial buffer sharing:* high priority cells can enter the buffer while the buffer is not full. Low priority cells can only enter the buffer if the total buffer occupancy is less than a threshold value. Obviously the threshold value is smaller than the total capacity of the buffer. The threshold level may be adjusted to suit different loadings.

- *Buffer separation:* separate buffers are used for low priority and high priority cells. This mechanism is not suitable for a connection that generates both priorities because in such situation the cell sequence integrity cannot be maintained.

Out of the three buffering mechanisms outlined in [53] partial buffer sharing is found the best compromise between system performance and buffer management complexity.

Bonomi et al. [54] analyse a partial buffer sharing system. The buffer is fed by burst-silence sources. The peak bit rate of a source is considered to be the same as the link bit rate. A Bernoulli trial is used to classify the arrivals into Class 1 (higher priority) and Class 2 (lower priority). Hou and Wong [55] outline a partial buffer sharing strategy for multiplexing burst-silence sources (high loss priority) and geometric arrivals (low loss priority). They assume that Class 1 arrivals are prior to Class 2 arrivals during any service interval. Yin et al. [56] use a fluid flow approximation of burst-silence sources for an approximate performance analysis of a partial buffer sharing strategy. Cheng and Wu [57] provide an approximate analysis for a generalised partial buffer sharing system. The buffer is fed by Poisson batch arrivals and deterministic service time is assumed. Cheng and Wu also outline and analyse an special case of a push-out scheme where replacement of the cells in the buffer is only possible for the arrivals within the current service interval. Finally, Lucantoni and Parekh [59] outline and analyse a partial buffer sharing system where the arrivals are Poisson and infinite buffer size is assumed.

A common finding in most of these studies is that priority control at the buffer can improve the system performance.

1.6.4 Congestion Control

Let us quote the definition of congestion in B-ISDN from CCITT [24]: *'In B-ISDN, congestion is defined as a state of network elements (e.g. switches, concentrators, transmission links) in which, due to traffic overload and/or control resource overload, the network is not able to guarantee the negotiated QOS to the already established connections and to the new connection requests'.* This definition implies that any buffer saturation cannot be regarded as congestion because depending on the type of service, it is possible to get buffer saturation and still meet the negotiated QOS. Congestion may be caused by a fault or failure in parts of the network. It can also be caused by unpredictable fluctuations in the traffic of various sources.

Congestion control refers to the set of actions taken by the network to stop the spread of congestion and to minimise its duration and its effect on the QOS. The two levels of traffic control outlined earlier, namely connection admission control and usage parameter control may be regarded as congestion control capabilities.

The factor that enhances the wasting of resources during congestion is that when users detect the loss of some of their cells, they start retransmitting and may even retransmit some cells that have successfully been delivered [60].

One method of tackling congestion control is to reduce the risk of overload to negligible levels [61] by ensuring that even during an overload condition, scarce resources are not a concern. This method does not make efficient use of network resources. The alternative is to design congestion control strategies that minimise the waste of network resources during congestion. This could be achieved by encouraging the users to adaptively control their offered load and reduce it during network overloads. In other words the network must create an incentive for users to use the available resources during congestion in a fair and responsible manner.

One approach to ensuring that all users get a fair share of network resources during congestion is *fairness queueing* proposed by Nagle [62]. When a network becomes congested, all the cells waiting for that network's resources are sorted

into separate queues, one queue per (source) user. An example of this is shown in Figure 1.17 [62]. Each queue has a finite size and those cells that encounter a full queue are discarded. The cells of each queue are served on a FIFO basis. The server switches over the queues in a round robin fashion. When the server visits a queue, say queue A, it serves one cell from that queue. It then switches over to another queue and does not visit queue A again until all other non-empty queues have been attended and one cell from each of them has been served. Empty queues are skipped over and loose their turn.

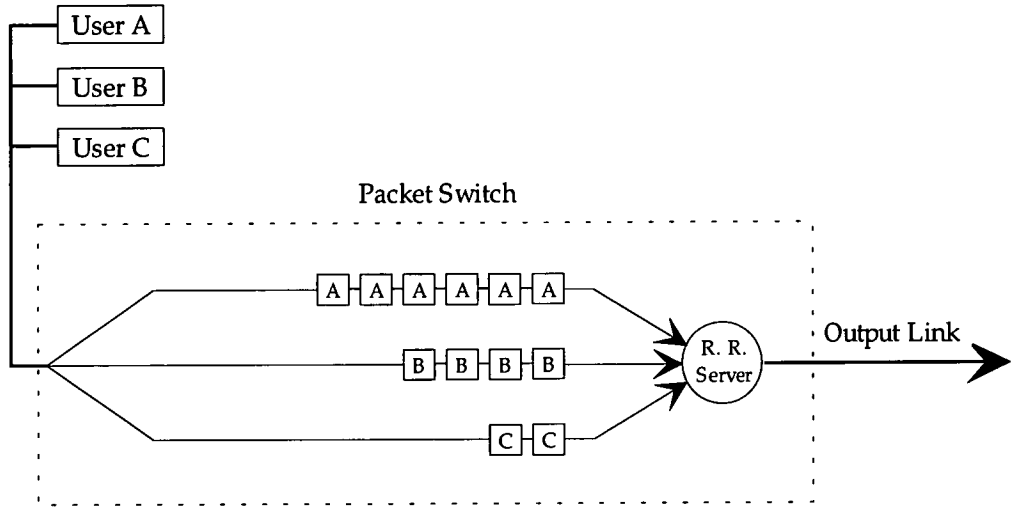


Figure 1.17: Fair queueing

One of the advantages of fairness queueing is that if a user attempts to send more cells than its fair share of the available capacity during an overload, only that user will be penalised by incurring a higher cell loss rate. Other users will be immune from the misbehaviour of that user.

Another congestion control method called *fairness discarding*, similar in concept to fairness queueing but more economical to implement, has been proposed by O'Neill [60]. Fairness discarding algorithm is shown in Figure 1.18 [60]. This algorithm is more economical to implement than the fairness queueing method because it does not require separate queues for each user. Only one queue is used for all users. The fairness discarding is implemented prior to the FCFS queue.

This means that in the event of network congestion, if a user sends more cells than his fair share of the available capacity, the violating cells are discarded before they enter the FCFS queue.

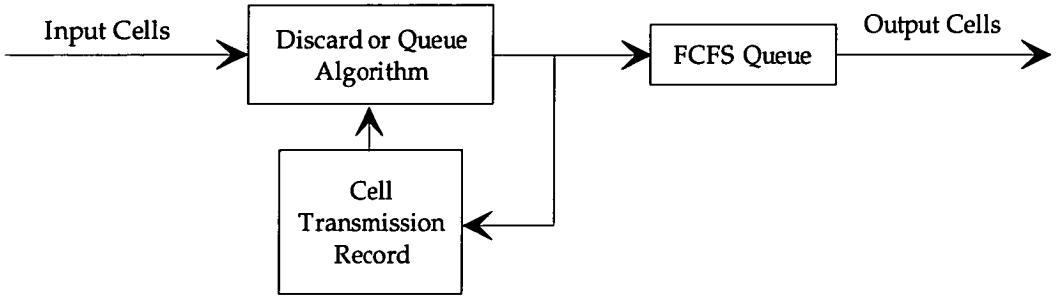


Figure 1.18: Fairness discarding

In order to implement the fairness discarding algorithm, some record of the cell transmission from each user is kept. The range of parameters that may be used for the discard algorithm, and the discard algorithm itself, offer a wide range of options for resource management during congestion.

Although congestion control in B-ISDN has been the subject of research for a number of years, definitive solutions for it are still lacking. It is not yet sufficiently clear how the dynamics of the network would behave in the presence of congestion, and how the traffic and the network resources should be handled to obtain more predictable and more reliable performance.

The uncertainties in the traffic patterns of various services and the time varying nature of networks dynamics and conditions suggest that some adaptation capability is desirable in the congestion control design of ATM networks. This explains some current interest in the applications of neural networks to ATM network design and control problems [63], such as the shortest-path routing [64], the admission control [65], and the traffic flow assignment [66].

1.7 The Contributions of this Thesis

The major thrust of this thesis is to study performance analysis models for Broadband Integrated Services Digital Networks. Chapters 2, 3 and 4 focus on several performance models that are applicable to access control problem.

In chapter 2, two access control strategies that may be suitable for those cases where two types of traffic are multiplexed are outlined. The two types of traffic are designated as Wideband (WB) and *Narrowband* (NB). The WB traffic may be assumed to be generated from Constant Bit Rate (CBR) services which are sensitive to delay and to delay variations and require uninterrupted, priority transmission once their connection is established. The NB traffic may be considered to be generated from data services (e.g. multiplexed traffic from several data sources) which are sensitive to packet loss, but insensitive to packet delay or the variations in the packet delay. The performances of these strategies are assessed using several techniques, namely simulation, an approximate Markov chain analysis (iterative), an iterative method of matrix-geometric analysis, and a decomposition method of matrix-geometric analysis. A comparison is made of the computational effort required for each approach.

Chapter 3 uses simulation tools to study the problem of mixing interactive data, interactive image, and video traffic at an ATM multiplexer. It shows that when all sources have similar burst bit rates and when the burst bit rates are much smaller than the link bit rate, then reasonably high utilisations may be achieved under very simple access control strategies. This study shows how the performance of the access node is affected as a result of reducing the ratio of multiplexer output link bit rate to the burst bit rates of the incoming traffic. This chapter also considers the effect of introducing priority to the cell stream of the video traffic on the performance of the access node.

Chapter 4 proposes and analyses an access control strategy for an ATM access node multiplexer that serves a mixture of CBR and VBR traffic. The performance parameters for the CBR traffic are obtained using a Markov chain. The

VBR traffic component is analysed using an imbedded Markov chain. All results are verified by simulation.

Because accurate source models are an integral part of performance analysis of a network, two chapters of this thesis have been dedicated to source traffic models as applicable to B-ISDNs. The main focus of chapters 5 and 6 are traffic models for video services in broadband networks. Special attention is paid to traffic models for video services because the wide range of video services and their relatively large bandwidth suggests that they will greatly affect the overall performance and bit rate requirements of B-ISDNs.

Chapter 5 provides a literature survey of the models proposed for video services. Chapter 6 investigates several models for video services, based on hidden Markov models. The performance of these models are evaluated. Some correlation analysis of the video traffic is also undertaken, indicating the presence of cyclic variations in the traffic generated from variable bit rate video codecs.

The cyclic variations in video traffic prompted an investigation into analytical models that can handle queueing systems with periodic variations in the arrival rate and/or service rate of their traffic. The results of these investigations are presented in chapters 7 and 8.

Chapter 7 begins with a M/M/1 queueing system and proceeds to the situation where the arrivals are cyclo-stationary and have a mean value that is sinusoidal in shape. A method of analysis is presented which uses Fourier series for calculating the cyclo-stationary probabilities of being in different states of the queueing system. Performance results are presented. The effects of truncation of the Fourier series on the accuracy of the results are investigated. The effect of the frequency of the cyclo-stationary arrival rate on the performance of the system is also investigated. The solution is then generalised for a queue that has a cyclo-stationary arrival rate with an arbitrary shape. Several examples are solved using the computational probability method presented in this chapter.

Chapter 8 considers a queueing system where both the arrival rate and the service rate of the queue have a cyclo-stationary nature and have sinusoidal shapes. The solution of chapter 7 is adapted for this situation and the conditions for which a numerical solution may be obtained are investigated. The solution technique is then extended to cater for arbitrary shapes of the arrival rate and the service rate.

Chapter 9 provides an overall summary of the performance models outlined in this thesis, and the major results obtained from them. It also suggests some extensions to this research for future work.

Chapter 2

Movable Boundaries in Dynamic Allocation of Capacity

2.1 Introduction

In this chapter, several access control strategies that may be applicable to both ATM networks and synchronous TDM networks, carrying a mixture of two types of traffic are studied. The traffic types are designated in this chapter as *Wideband* (WB), identified loosely with video, and *Narrowband* (NB), typifying data traffic. The WB connections are assumed to be intolerant of delay or delay variation, and to require uninterrupted, priority transmission once established. WB connections are either accepted, in which case the required transmission capacity is allocated for the whole duration of the call, or else blocked. The NB traffic is taken to be less time-critical but loss sensitive. NB connections are not blocked, but may be delayed or queued if transmission capacity is temporarily unavailable.

The control strategies investigated partition the available transmission capacity into WB and NB allocations. There is a minimum guaranteed capacity available for the NB traffic. Under some conditions the boundary between the NB and WB allocations can move, and in one model there is provision for pre-emption of low-priority NB calls. The performance of a system using under *Movable Boundary* strategy is assessed. The performance parameters considered
are the blocking probability of WB calls and NB call delay.

The two strategies differ in their assumptions about the mode of operation of the traffic multiplexer. The first model is consistent with the lower-priority traffic being delivered to the multiplexing point on NB lines, with limited bit rates and flow control. This system has strong parallels with an approach that has been used in TDM studies. The second model is more consistent with a conventional statistical multiplexing arrangement, with no explicit limits imposed by the bit rate at which incoming NB traffic might be delivered.

The performances of these strategies have been assessed using simulation techniques, by an approximate Markov chain analysis, and using matrix-geometric techniques. A comparison is made of the computational effort required by each method. The background theory required for the analyses presented in this chapter is given in Appendix A.

2.2 Related Work

Historically, there are several streams that have contributed ideas to the design of access control strategies:

- General interest in the classical problems of congestion control, grade of service and efficient link utilisation, which were originally studied in circuit-switched telephone systems.
- The realization that the separate networks which historically evolved for separate services (e.g. telephony and data) are economically inefficient, and that a single integrated network supporting a variety of services would be better. This concept (effectively ISDN) implies a mix of traffic types competing for shared network resources, with a variety of performance criteria.
- The development of packet switched networks. In its original form, the conceptual packet switching network would have treated all packets in the same manner, without differentiating between traffic types or supporting different priority levels. This uniform processing of all packets requires that

a single grade of service be adequate for the most demanding services, and is inconsistent with efficiently meeting a range of different performance targets for different traffic types. These basic difficulties in coping efficiently with different traffic types while maximising the link utilisation were recognized by Kleinrock [67].

As a particular case, problems could arise if due to the onset of congestion in the network, packets with network control messages were unable to use a higher priority to reach their destinations reliably and promptly.

- Investigations of the cost and performance of hybrid networks combining circuit-switched and packet-switched arrangements to accommodate mixed traffic types. Generally each traffic class is handled by the transmission technique which can best satisfy the performance criteria for the class.

- The adoption of ATM by CCITT as the proposed mechanism for implementation of Broadband ISDN systems [23].

In resolving the conflicting requirements of high link utilisation and acceptable performance for each traffic type, a number of integrated control strategies have been proposed that aim at a controlled balance between these two objectives. Some of these are strictly applicable to systems in which capacity allocation is handled by assigning sufficient channels (timeslots) to a call to accommodate its bit rate, as in a TDM system. Some other strategies mentioned in the literature are applicable to statistical multiplexing environments, in which the full unallocated capacity left after satisfying priority customers can be used to rapidly drain queues of low-priority traffic. This is more obviously relevant to the regime in which an ATM multiplexer will operate. For systems involving mixtures of more than two traffic types with different priorities and performance criteria, elements of both approaches are likely to contribute to the analysis of access control algorithms, and both approaches are covered here. The two formulations will be referred to as the *TDM* approach and the *ATM* approach.

A typical full analysis of the integrated access control problem can be set up as a multi-server arrangement, in which each identical server corresponds to one

timeslot in a TDM frame on the outgoing digital pipe. A queue can be established for each traffic type, with overflow between the queues under some conditions. Unfortunately, this model turns out to be intractable for more than two traffic types, due mainly to the dimension of the state space which results [68, 69]. Most subsequent analyses have been concerned with a small number of traffic types, usually two, generally chosen to represent traffic classes with the two main types of performance criteria: blocking and delay, and often with one "wideband" and one "narrowband" type.

Several examples exist in the literature of schemes in which each of the traffic types has equal access to the whole available transmission capacity of the outgoing link. This is called "complete sharing". The case of queueable wideband and blockable narrowband traffic is treated by Gimpelson [70], two heterogeneous classes identified with voice and data by Bhat [71], and an infinite system of blockable wideband and queueable narrowband traffic by Kraimeche [72]. This general approach is attractively simple, but can suffer from bandwidth inefficiencies if, for example, traffic of one class has to wait behind a message of another class whose transmission requirements take a while to become available.

At the other extreme, each available TDM channels can be allocated for the exclusive use of a particular traffic class. This is called "complete partitioning". This strategy, sometimes known as a "fixed boundary" system, has been studied for a mixture of circuit-switched and packet switched traffic by Ross [73]. Obviously such a strategy prevents interaction between the two types of traffic, and the grades of service for each individual type can be improved. The main disadvantage of such scheme is that under unbalanced load conditions the link utilisation is poor, and some unused channels cannot be put into service to improve the performance of the more active traffic type.

The desirable features of complete sharing and complete partitioning may be combined to some extent by allowing dynamic reallocation of some channels. This is the "Movable Boundary" scheme, and has been investigated for combinations of circuit-switched voice and packet-switched data by Kummerle [74], Coviello et

al. [75] and Fischer [76].

Consider the case of two traffic types with significantly different bandwidth requirements. Assume that narrowband calls are allowed to occupy channels accessible to wideband traffic. If servicing of new wideband calls requires the allocation of a contiguous block of channels, then inefficiencies can result when one or two NB channels effectively prevent access to many channels for the WB traffic. A remedy for this problem, proposed by Yamaguchi and Akiyama [77], is to allow call pre-emption such that the WB calls can pre-empt NB calls occupying WB channels. The performance of this type of algorithm for voice and data traffic has been analysed by several authors [78, 79, 80, 81, 82, 73, 83].

Indiscriminate use of call pre-emption can however result in long delays for the pre-empted NB calls if there is heavy wideband traffic. By using a "restricted access" scheme in which there is some minimum allocation of channels guaranteed for the NB traffic (and by implication, a maximum allocation enforced for the WB calls), this problem can be alleviated. This is the situation covered in Kraimeche [72] for circuit-switched calls, and by DeSerres and Mason [84] for a mixture of blockable narrowband and queueable wideband types.

In his thesis, Chua [85] has extended such scheme to allow the boundary between the channel allocations to be moved in *either* direction so that there are limits on the maximum capacity allocated to each type, with a range of dynamic control available between these limits. This strategy partially overcomes the inefficiency that may result when the WB traffic reaches its limit in a movable boundary scheme, but narrowband traffic load is light. Chua has also generalized this system further by providing for queueing of both types of traffic in a "buffered bidirectional movable boundary" strategy, with pre-emption under some conditions [85].

Questions of channel assignment disappear, and pre-emption is effected automatically in some models when the ATM approach is used. Examples are given in [86] for subscriber links and in

[87], [88] for the B-ISDN network, in which two broad

classes of traffic (namely time-critical, and queueable) are handled and the low priority bursty traffic is able to use whatever capacity is left by the high priority virtual circuit traffic. These studies define a capacity boundary, setting a minimum allocation for the NB traffic and a corresponding maximum for the WB traffic. The NB queue is essentially served at a fluctuating rate which may range from the full outgoing link capacity down to the minimum NB allocation, depending on the WB activity. Thus a movable boundary is in effect.

The analyses carried out by Chua [85], in which channel assignment is the underlying mechanism (the TDM approach), and [86] and [88] which essentially exploit statistical multiplexing advantages (the ATM approach) are both based on the matrix-geometric method developed by Neuts [89]. These studies, with common mathematical elements and similar purposes, form a starting point for our investigations.

2.3 Access Strategies for Non-Statistical TDM Multiplexer

In this section we consider a non-statistical TDM multiplexer which serves two types of traffic. Assume that the capacity of the digital pipe (the output link of the multiplexer) is comprised of C channels. The traffic in the network is divided into two general groups: narrowband (NB) traffic and wideband (WB) traffic. It is assumed that the inter-arrival time for calls of both NB traffic and WB traffic are exponentially distributed with means of $1/\lambda_1$ and $1/\lambda_2$ and that the service time distributions are exponential with mean service rates of μ_1 and μ_2 respectively. The NB and WB calls occupy b_1 and b_2 channels respectively. Let N1 and N2 denote the maximum number of NB and WB calls which can be in service. With these preliminary definitions we can now consider the individual control strategies in the following subsections.

2.3.1 Movable Boundary with no Sorting of Channel Allocations of the Digital Pipe (MBNSD):

This strategy is formulated as following:

- There is a guaranteed capacity allocated for m_1 NB calls.
- The remaining capacity of the link, R ($R = C - m_1 b_1$), can be shared by both types of traffic. When there are no NB calls placed on R, the shared capacity can carry a maximum of m_2 WB calls.
- Any additional NB call in excess of m_1, will be allocated channels from the shared capacity of the link, provided that capacity is not being utilised by WB calls. Therefore $N1 = m_1 + \lfloor m_2 b_2 / b_1 \rfloor$ where $\lfloor x \rfloor$ denotes the largest integer in $(0, x)$. When there is no capacity left in the link for an additional NB call, any further NB calls will be placed in a queue.
- A NB call that has been allocated from the shared capacity will continue using that capacity until it ends transmission, even if capacity becomes available in the guaranteed portion for NB calls. This is where MBNSD strategy is different from MB strategy which will be discussed later.
- WB calls can only transmit on R, therefore $N2 = R/b_2 = m_2$. If there is no capacity left from R for an additional WB call, any further WB calls will be lost.
- Unlimited buffer space is assumed to be available for queueing the non-priority traffic.

Note that under this strategy a situation could arise where there is enough capacity in the whole frame for a new WB call, but a WB call is denied access because the available capacity is not in the form of consecutive time slots. In other words it is assumed that access controller is not able to dynamically reassign capacity to calls which are in progress.

Because the capacity reallocation is not automatic or not possible under this strategy, the transition probabilities leading from a particular state depend not

only on the number of NB and WB calls in that state, but also on how the current number of NB calls were allocated capacity, i.e. on the history of the process. In short, a 2-D Markov analysis is not possible.

Analysis of MBNSD Using Simulation

This case has been assessed only by simulation methods, but a three-dimensional formulation of this problem can be set up by identifying and including in the state specification three call types: NB using NB channels, NB using WB channels, and WB calls. The simulation results are given, along with the results of other strategies, in section 2.5, Figures 2.5 to 2.7. The advantage of simulation methods is that almost any traffic pattern can be modelled by altering the random number generating functions within the simulation program. Even when a mathematical model is developed, simulation may be used as a versatile tool to verify and check the accuracy of the analytical model. However, the drawback of using simulation is that a lot of computer time is required to obtain statistically reliable results. To simulate these strategies, the SIMSCRIPT-II.5 package was used. This package is preferable to general purpose programming languages because Simscript is very English-like and self documenting, produces very efficient code, and reduces the time spent on writing the code significantly.

Almost all the simulation languages use one of the two basic approaches to simulation modelling, here referred to as the event scheduling approach and the process approach. This work uses the process approach. Each call is treated as a process that enters the system, stays in the system for some time and then leaves. The simulated time is chosen so that there are enough arrivals of the least probable event to guarantee an acceptable accuracy for the output results. This is done by running the simulation program several times with the same set of input data, but with different simulated times. Each time the number of the least probable event is increased and the output results are compared with the previous sets. This process is repeated until the variation in the output results is less than the degree of accuracy required.

2.3.2 Movable Boundary with Sorting of Channel Allocations of the Digital Pipe(MB):

In addition to the formulation for MBNSD, MB assumes that when a NB call transmitting over the guaranteed capacity ends transmission, if there are no other NB calls in the queue and if there is a NB call transmitting over the shared capacity, that will be reallocated capacity from the guaranteed portion of the NB calls. This would depend on the access control ability to dynamically reassign the channels used by an individual call without significant interruption to the call. If it can do so without large overheads, then NB calls using WB channels can be reassigned to NB channels which become free due to service completion of a NB call. This would keep the two traffic types apart as much as possible and would minimize interaction between them. If this dynamic reallocation of calls to channels is not possible then NB calls may well be left occupying WB channels, even though many NB channels are free.

In terms of the model, there is a considerable simplification which results in the case where dynamic reallocation is possible. In particular, it allows the state (in the Markov sense) of the system to be specified fully by the number of calls of each type currently in progress, thus leading to a 2-D Markov model.

The allocation of available capacity is illustrated in Figure 2.1. The limits imposed by the outgoing link capacity are shown as a smooth line but in reality this limit will be a stepped or "staircase" boundary. The upper limit on the number of simultaneous WB calls has been shown as WB_{max}. This is because of the guaranteed minimum capacity for the NB traffic. There are three patterns of service for NB calls, indicated by the regions labelled 1, 2 and 3 in Figure 2.1:

- In region 1, NB calls occupy only NB channels.
- In region 2, more NB calls are in progress. All the channels designated NB are in use, and NB calls are additionally using some of the WB allocation. From region 2, the link capacity limit cannot be crossed due to connection of another WB call and hence blocking of WB calls will result.

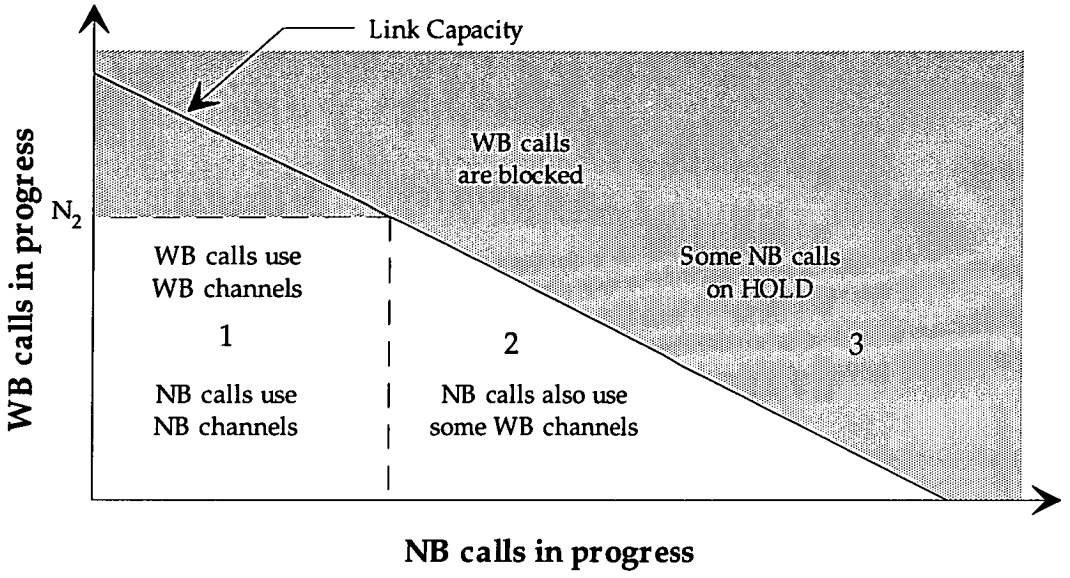


Figure 2.1: Capacity allocation for the MB strategy

- In region 3, accessible only by connection of large numbers of NB calls, the number of WB calls can only remain the same or decrease.

Analysis of MB using a Finite-State Markov Chain

Here the state of the system is defined by (i, j) where i and j give the number of NB and WB calls in the system respectively. The constraints on the motion of the system state are reflected in the allowed transitions and their probabilities shown in Figure 2.2. Note that again the channel capacity limit is shown as a smooth boundary. Except for those states on the boundaries $i = 0$, $j = 0$, $j = \mathrm{WB}_{max} = N_2$, the probabilities shown reflect the four possible state transitions which may occur:

1. Arrival of another WB call (λ_2)
2. Completion of one of the j WB calls in progress ($j\mu_2$)
3. Arrival of another NB call (λ_1)
4. Completion of service for a NB call. Note that for states to the left of the channel capacity limit in Figure 2.2, all i NB calls in progress are being

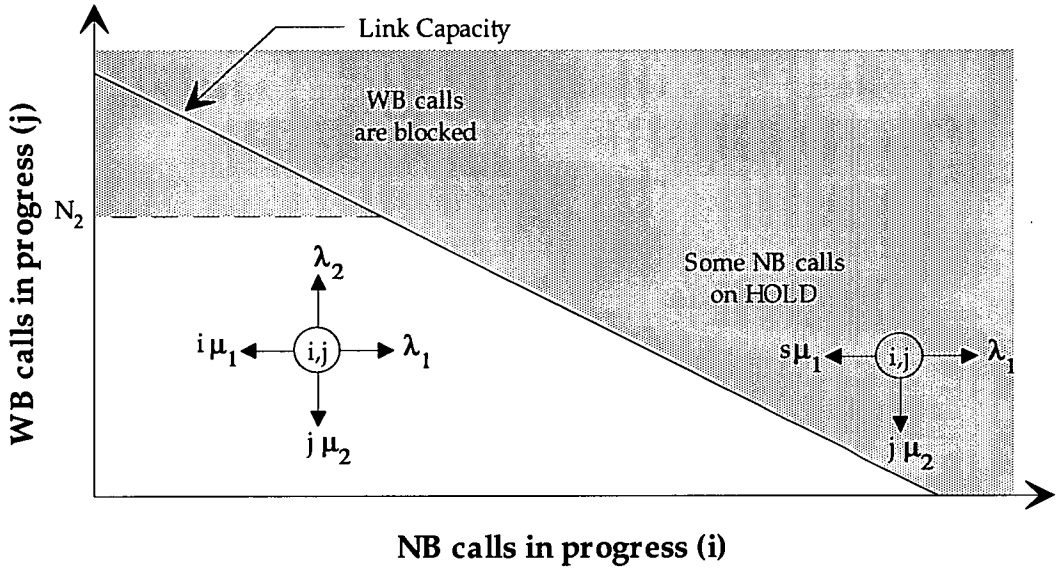


Figure 2.2: MB Transition Probabilities

served, yielding a probability $i\mu_2$. For states to the right of the channel capacity limit, the number of WB calls in progress determines the capacity currently available for NB calls. Of the i NB calls in progress, only s ($s < i$) are receiving service and are candidates for completion. The corresponding probability entry is $s\mu_1$.

For the purpose of this analysis, a limit (l) is placed on the queue length of the NB calls to limit the number of Markov states, i.e. $l = \mathrm{NB}_max$. This limit must be selected such that the probability of the NB queue being greater than that limit is very small. The results obtained for this truncated system then will be a good approximation to the behaviour of the actual system. This approach is only valid when both WB and NB traffic are Poisson, with exponentially distributed service times. The computation of the MB stationary probabilities entails

1. Assigning a state number n; $n = 1, 2, \ldots, M$ to the states (i,j).

2. Constructing the $M \times M$ transition rate matrix Q.

3. Solving for the stationary probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_M)$ given by

$$\pi Q = \mathbf{0}$$

and

$$\sum_{n=1}^{M} \pi_n = 1$$

determines the probability of each state.

4. For each such state n, the number of WB and NB calls in progress and the number of NB calls on "hold" (if any) is known. The probabilities π_n, $n = 1, 2, \ldots, M$ are then used to determine the distributions of WB and NB calls in progress. The tail of the distribution for NB calls indicates whether NB_{max} has been chosen sufficiently large.

5. The blocking probability P_b for WB calls follows from the arrival rate λ_2 and the average WB traffic actually carried.

6. The average number of calls on hold follows from π_n and the $s = s(n)$ values. Little's result [90] is used to convert this answer to an average queueing delay time for NB traffic.

In solving the potentially large set of simultaneous linear equations for the stationary probability vector π, an iterative method has been chosen in the interests of numerical accuracy. The specific algorithm currently implemented is a Gauss-Seidel iterative scheme using successive over-relaxation, a system which gives reliable and accurate convergence, but at some cost in speed.

Analysis of MB using Matrix-Geometric Methods

In this subsection the performance parameters of the non-statistical TDM multiplexer under MB access control strategy are obtained by the matrix-geometric solution techniques. This analysis is as presented in [85] and is given here to show an alternative to other solution methods that have been used throughout this chapter.

Under MB strategy, the access controller node can be modelled as a two-dimensional Markov process. The state of the system may be defined by (i, j) where i and j denote the number of NB and WB calls in the system respectively.

In order to quantify the performance measures we need to obtain the stationary probabilities of different states (i, j) of the system, i.e. for all allowed values of i and j we need to find p_{ij} where

$$p_{ij} = \text{Prob}(i, j) \ .$$

Let Q'' denote the transition rate matrix of the Markov process. Furthermore, let $\mathbf{P_i}$ be the stationary probability vector for a particular number of NB calls in the system, i.e.

$$\mathbf{p_i} = [p_{i0}, p_{i1}, ...] \ .$$

The stationary probabilities for all states can then be written as $\mathbf{p}$ where

$$\mathbf{p} = [\mathbf{p_0}, \mathbf{p_1}, \mathbf{p_2}, ...] \ .$$

When the process is positive recurrent, this vector may be calculated by solving for the following overdetermined set of linear equations:

$$\mathbf{p}Q'' = \mathbf{0}$$

$$\mathbf{p}E = 1 \ . \tag{2.1}$$

Here, E is a vector with all its elements being column vectors. All entries of these column vectors are 1. The state transition matrix, Q'', for this system may be written as:

$$Q'' = \begin{bmatrix} A_{01} & A_{01} & & & & & & \\ A_{12} & A_{11} & A_{10} & & & & & \\ & A_{22} & A_{21} & A_{20} & & & & \\ & & . & . & . & & & \\ & & & A_{M-1,2} & A_{M-1,1} & A_{M-1,0} & & \\ & & & & A_2 & A_1 & A_0 & \\ & & & & & . & . & . \end{bmatrix} \tag{2.2}$$

where M is an index to the level of states separating the boundary states from the non-boundary states. The structure of Q'' is similar to that of the quasi-birth-death process (see Appendix A). The entries of matrix Q'' as given by equation

2.2 are themselves square matrices which contain the various arrival and service rates of the two traffic types. These matrices are of the size $(N-2+1)\times(N_2+1)$ and under this access strategy they can be defined for $k=0,1,2,\cdots,M$ as

$$A_{k0}(i,j) = \begin{cases} \lambda_1 & , \quad i=j \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$A_{k1}(i,j) = \begin{cases} \lambda_2 & , \quad i=j-1, 1 \le m_2 - \lceil (k-m_1)b_1/b_2 \rceil - i \\ i\mu_2 & , \quad i=j+1 \\ a_k(i) & , \quad i=j \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$A_{k2}(i,j) = \begin{cases} \min(k, m_1+\alpha_i)\mu_1 & , \quad i=j \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $\lceil x \rceil$ is the smallest non-negative integer greater than or equal to x,

$$a_k(i) = -\{A_{k0}(i,i) + A_{k1}(i,i-1) + A_{k1}(i,i+1) + A_{k2}(i,i)\} \quad ; \quad 0 \le i \le N_2$$

and

$$\alpha_j = \lfloor (N_2 - j)b_2/b_1 \rfloor \ .$$

The following theorem[91] is required for our analysis (proof omitted):

Theorem 1 *When the Markov process Q'' is positive recurrent, the steady-state probability vector $\mathbf{p} = [\mathbf{p_0}, \mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_{M-1}}, \mathbf{p_M}, ...]$ is given by*

$$\mathbf{p_i} = \mathbf{p_{M-1}} R^{i-M+1} \quad , \ i \ge M \ . \tag{2.3}$$

The matrix R is the unique solution of the matrix-quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0 \ . \tag{2.4}$$

The probability vector $\bar{p} = [p_0, p_1, p_2, ..., p_{M-1}]$ is the unique solution of the equations

$$\bar{\mathbf{p}} T = 0$$

$$\bar{\mathbf{p}} E + \sum_{i=M}^{\infty} \mathbf{p_i} e = 1 \tag{2.5}$$

where the matrix T is obtained by truncating Q'' to that corresponding to $\bar{\mathbf{p}}$ in the general equation $\mathbf{p}Q'' = 0$. Therefore

$$T = \begin{bmatrix} A_{01} & A_{01} & & & & & \\ A_{12} & A_{11} & A_{10} & & & & \\ & A_{22} & A_{21} & A_{20} & & & \\ & & \cdot & \cdot & \cdot & \cdot & \\ & & & A_{M-2,2} & A_{M-2,1} & A_{M-2,0} & \\ & & & & A_{M-1,2} & A_{M-1,1} + RA_2 \end{bmatrix} . \tag{2.6}$$

The solution method provided by Neuts[89] involves the following steps:

- The stability of the system is checked by testing for positive recurrence of the matrix Q''.
- If Q'' is positive recurrent, the matrix R is then determined by applying equation (2.4) recursively, beginning with the estimate $R(0) = 0$.
- Finally the invariant probability vector $\bar{\mathbf{p}}$ is obtained from equation (2.5).

When the dimensions of Q'' are large, there may be computational difficulties associated with the last step of the process. Hence, direct matrix inversion may not provide results with the desired degree of accuracy. A technique for reducing the dimension of Q'' has been introduced in [81] but in this work a method of decomposition introduced in [92] has been used.

In order to find the stationary probability vector $\bar{\mathbf{p}}$ let us first rewrite equation (2.5) as

$$\mathbf{p_0}A_{01} + \mathbf{P_1}A_{12} = 0$$

$$\mathbf{p_{i-1}}A_{i-1,0} + \mathbf{p_i}A_{i,1} + \mathbf{p_{i+1}}A_{i+1,2} = 0 \quad , \quad 1 \leq i \leq M-1$$

$$\mathbf{p_{M-2}}A_{M-2,0} + \mathbf{p_{M-1}}(A_{M-1,1} + RA_2) = 0 \ . \tag{2.7}$$

Equation (2.7) implies that

$$\mathbf{p_i} = \mathbf{p_{M-1}}H_i \quad , \quad 0 \leq i \leq M-2 \tag{2.8}$$

where matrices H_i may be calculated recursively from:

$$\begin{aligned} H_{M-1} &= I \\ H_{M-2} &= -(A_{M-1,1} + RA_2)A_{M-2,0}^{-1} \\ H_{M-i} &= -(H_{M-i+1}A_{M-i+1,1} + H_{M-i+2}A_{M-i+2,2})A_{M-i,0}^{-1} \quad , \quad 3 \leq i \leq M. \end{aligned} \tag{2.9}$$

Finally, from the first equation in (2.7) and the normalising equation in (2.5), $\mathbf{p_{M-1}}$ can be calculated as:

$$\mathbf{p_{M-1}}(H_0A_{01} + H_1A_{12}) = 0$$

$$\mathbf{p_{M-1}}\{\sum_{i=0}^{M-2} H_ie + (I-R)^{-1}e\} = 1 \ . \tag{2.10}$$

These give an overdetermined linear system of equations which can be solved using standard mathematical library routines. As far as the stability of the system is concerned, it can be shown that [89] the process Q'' will be positive recurrent if and only if the matrix R has all its eigenvalues inside the unit disc, i.e. has spectral radius $sp(R) \leq 1$. This holds when the matrix $A = A_0 + A_1 + A_2$ is irreducible which is the case if and only if

$$\chi A_2 e > \chi A_0 e \tag{2.11}$$

where χ is the stationary probability vector of A. Since the matrix A is irreducible, Q'' is positive recurrent if and only if

$$\sum_{i=0}^{N_2} \chi_i A_2(i,i) > \sum_{i=0}^{N_2} \chi_i \lambda_1$$

or

$$\sum_{i=0}^{N_2} \chi_i (A_2(i,i) - \lambda_1) > 0 \ . \tag{2.12}$$

Because $\chi_i \geq 0$ for all i, positive recurrence of Q'' is guaranteed when

$$A_2(i,i) \geq \lambda_1 \quad , \quad \text{for all i} \ . \tag{2.13}$$

In particular, since $A_2(i,i)$ is minimum when $i = N_2$, positive recurrence of Q'' is guaranteed when

$$m_1\mu_1 > \lambda_1 \ . \tag{2.14}$$

This equation simply implies that so long as the service rate for NB traffic exceeds its arrival rate, the system will be stable. However, because under this strategy NB calls can also transmit on unused WB channels, the stability condition given by (2.14) is sufficient but not necessary.

Let $E[n_1]$ be the average number of NB calls in the queue. This can be calculated as

$$E[n_1] = \sum_{j=0}^{N_2} \sum_{i=m_1+\alpha_j}^{N_1-1} (i - m_1 - \alpha_j)p_{ij} + \sum_{i=N_1}^{\infty} \sum_{j=0}^{N_2} (i - m_1 - \alpha_j)p_{ij} \ . \tag{2.15}$$

Let α be the vector $[\alpha_0, \alpha_1, ..., \alpha_{N_2}]$. Then by (2.3) and noting that $sp(R) < 1$, equation (2.15) can be rewritten as

$$\begin{aligned} E[N_1] &= \sum_{j=0}^{N_2} \sum_{i=m_1+\alpha_j}^{N_1-1} (i - m_1 - \alpha_j)p_{ij} + p_{N_1-1}R(I-R)^{-1}\{(N_1 - m_1)e - \alpha\} \\ &+ p_{N_1-1}R^2(I-R)^{-2}e \ . \end{aligned} \tag{2.16}$$

Now, we can use Little's result [93] to calculate the average NB queueing delay from the ratio $E[n_1]/\lambda_1$. The variance of the NB queue may be given in terms of the first and the second moments of the NB queue size:

$$\mathrm{VAR}[n_1] = E[n_1^2] - E[n_1]^2 \tag{2.17}$$

where

$$E[n_1^2] = \sum_{j=0}^{N_2} \sum_{i=m_1+\alpha_j}^{N_1-1} (i - m_1 - \alpha_j)^2 p_{ij} + \sum_{i=N_1}^{\infty} \sum_{j=0}^{N_2} (i - m_1 - \alpha_j)^2 p_{ij} \tag{2.18}$$

or

$$\begin{aligned} E[n_1^2] &= \sum_{j=0}^{N_2} \sum_{i=m_1+\alpha_j}^{N_1-1} (i - m_1 - \alpha_j)^2 p_{ij} + p_{N_1-1}R\{(I+R)(I-R)^{-3} + \\ &\quad 2(N_1 - m_1 - 1)R(I-R)^{-2} + [(N_1 - m_1)^2 - 1)](I-R)^{-1}\}e \\ &+ 2p_{N_1-1}R\{(N_1 - m_1)(I-R)^{-1} + R(I-R)^{-2}\}\alpha \\ &+ p_{N_1-1}R(I-R)^{-1}\alpha^{(2)} \end{aligned} \tag{2.19}$$

where $\alpha^{(2)}$ is the vector $[\alpha_0^2, \alpha_1^2, ..., \alpha_{N_2}^2]$.

The only performance measure for the WB traffic is its blocking probability which can be calculated by:

$$\begin{aligned} PB_2 &= \sum_{i=0}^{m_1} p_i N_2 + \sum_{i=m_1+1}^{N_1-1} \sum_{j=N_2-\lfloor (i-m_1)b_1/b_2 \rfloor}^{N_2} p_{ij} + \sum_{i=N_1}^{\infty} \sum_{j=0}^{N_2} p_{ij} \\ &= \sum_{i=0}^{m_1} p_i N_2 + \sum_{i=m_1+1}^{N_1-1} \sum_{j=N_2-\lfloor (i-m_1)b_1/b_2 \rfloor}^{N_2} p_{ij} + p_{N_1-1} R(I-R)^{-1} e \ . \end{aligned} \tag{2.20}$$

The results of this method will appear in section 2.5.

2.3.3 Movable Boundary with Pre-emption(MBP)

The only difference between this strategy and the MB strategy is that WB traffic also has a guaranteed capacity for m_2 calls where $m_2 = R/b_2$. This means that a NB call transmitting over R will be pre-empted on the arrival of a WB call which can only transmit if it uses some of the capacity used by the NB call. Figures 2.3 shows the capacity allocation for this strategy. The limits imposed by the outgoing link capacity are again shown as a smooth line rather than a stepped boundary. By comparison with Figure 2.1 for the MB case, the modified conditions under which WB calls are blocked may be noted.

Analysis of MBP using a Finite-State Markov Chain

Here, as for the MB case, the state of the system is defined by (i, j) where i and j give the number of NB and WB calls in the system respectively. The constraints on the motion of the system state are reflected in the allowed transitions and their probabilities shown in Figure 2.4. As before, s represents the number of NB calls actually receiving service. Apart from these relatively minor differences, the MBP analysis follows closely that of the MB case given earlier, and the same program steps are involved. The same caveats on the number of states, NB_{max}, and the potential numerical difficulties also apply.

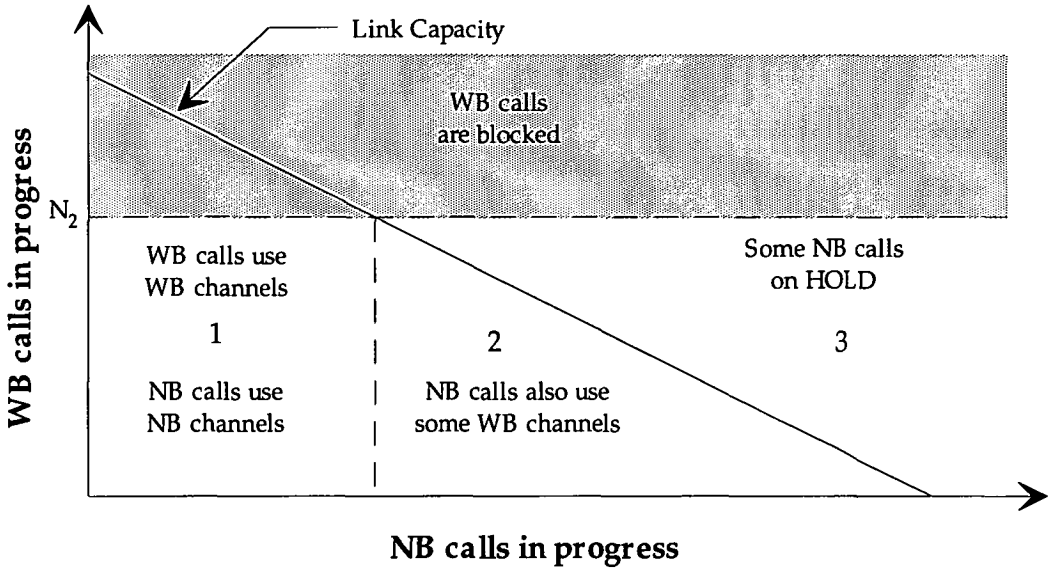


Figure 2.3: Capacity allocation for the MBP strategy

2.4 Access Strategies for Statistical ATM Multiplexer

This section defines the models used for the analysis of admission control in a system using the *ATM* approach. The general features of the problem are again restated within the specifications of each model. The first strategy considered in this section is very similar to MBP for the non-statistical TDM multiplexer given earlier and is analysed by an approximate Markov chain. The second strategy is again implementing MBP for an ATM multiplexer, but the problem set-up is more general and can handle multiple classes within the WB category with different arrival and service rates. Hence for the rest of this chapter we have combined the modelling assumptions, the strategy definition, and the analysis in the same subsection.

2.4.1 MBP and Markov Chain Analysis

This strategy is very similar to that described for the TDM non-statistical multiplexer. The main difference is that even is there are only a few NB calls in progress, all the available capacity of the link will be assigned to them. Such

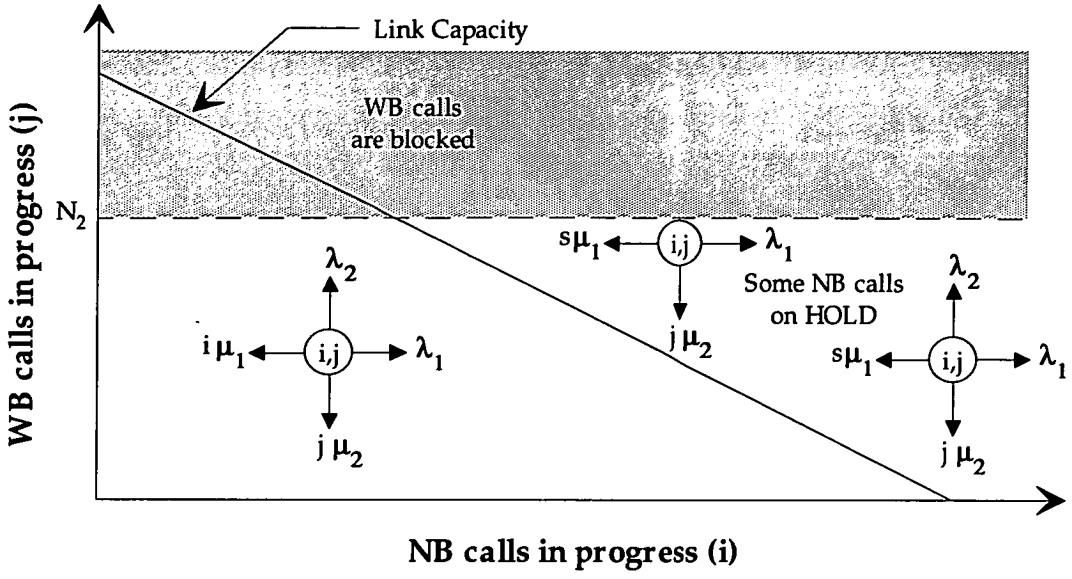


Figure 2.4: MB Transition Probabilities

messages would not have to be transmitted at a fixed, predetermined rate. Details of the model and the strategy are as follows:

- There are two types of traffic, designated WB and NB. Loosely, these are identified with video and data traffic. The WB traffic is time-critical and may not be queued, and a service request may be refused (blocked). NB calls are always connected, but they may be queued. No explicit limits are applied for the rate at which an input NB message becomes available; entry to the queue is assumed to be on a whole message basis[1].

- The outgoing line capacity is considered to be subdivided into WB and NB allocations. The WB priority traffic is to be handled as in the TDM approach, by allocating sufficient capacity for the whole duration of the call. It follows that some cyclic pattern of service will be required for the priority traffic lines, effectively establishing a cyclic pattern of cell use, with some designated as WB (priority), and the remainder as NB.

- The WB traffic is restricted to occupying only WB capacity. The NB capacity allocation guarantees a minimum capacity for NB calls, and any unused

[1]This implies local queueing, *cf.* the TDM approach.

WB capacity can be used by NB calls until it is needed for the transmission of a WB call.

- WB traffic may be blocked when a WB service request is received and all the capacity designated as "WB priority" has been engaged by other WB calls. The NB traffic is not blocked but may have to queue for service.

- The WB call holding times are taken to be exponentially distributed. The message length of a NB call is also taken to be exponentially distributed, but given the fluctuating service rate for the NB queue, there will be a different distribution of service time once the NB message has reached the head of the queue. Essentially, NB calls proceed at a transmission rate which is a function of the link load.

- If no limit is specified for the queueing space or line tables available for NB traffic, the number of NB calls in progress could in principle become very large. In this analysis, only a finite number of states are considered; this leads to a simpler formulation and reduced computation time. Strictly speaking, this model accurately describes a system in which only a finite number of NB messages can be handled, and when this limit is reached, blocking of NB calls would occur. In practice, by making the number of states large, the probability of NB blocking is made small and the results give a good description of the performance of a non-blocking system.

The computation of stationary probabilities for this approach involves similar steps to those given for the non-statistical TDM multiplexer.

2.4.2 MBP & Matrix Geometric Analysis

In this subsection the performance parameters of the non-statistical TDM multiplexer under a MBP access control strategy are obtained by matrix-geometric solution techniques.

Non-priority data traffic (NB) is taken in this instance to come from a switched Poisson process. Let λ_i denote the average arrival rate of calls in priority class i. These calls have an average duration $1/\mu_i$. This data may come from a number

of individual sources: here the aggregate data stream is modelled as a switched Poisson process (SPP) with two modes, denoted by $m = 0, 1$. In mode m, non-priority (NB) messages arrive according to a Poisson process with rate η_m. Mode changes occur at random; the mode duration parameters α_m determine the average time $1/\alpha_m$ spent in mode m.

We take the lengths of the non-priority messages to be exponentially distributed, with mean length L_{NB}. The traffic intensities and mode duration parameters of the SPP can be adjusted to reflect the "burstiness" of the non-priority data [94]. We now show how Neuts' method for queues in a random environment [89] can be used to reduce the system to an M/M/1 queue in a Markovian environment. Such systems are characterised by customer arrivals according to a Poisson process with average rate λ_j and an exponentially distributed service time with rate μ_j, where both λ_j and μ_j depend on the state j of the environment described by an irreducible Markov process.

The assumptions made imply that the arrival rate of non-priority data packets depends only on the SPP mode, while the fluctuating service rate for this NB traffic depends only on the number and class of priority customers currently connected. It is thus sufficient to define the "state" of the system by the number of priority connections in each class, and the mode of the SPP.

Suppose there are K classes of priority customers with different average call durations and/or line capacity requirements. The state of the system and the outgoing line capacity available for the non-priority traffic are both specified by a $(K+1)$–dimensional vector $S = [m, p_1, p_2, \ldots, p_K]$ embodying the modes $m = 0, 1$ of the SPP and the number p_i of priority calls of class i in progress. The state S also determines the transition probabilities to other accessible states, thus generating a Markov process.

The number of states M which S can attain is limited by the total line capacity available for priority traffic. Take the outgoing line capacity to be C, of which a maximum capacity C_p can be allocated to priority calls. If $C_p < C$ then there

is some guaranteed minimum throughput for the non-priority traffic. m can be found by enumerating all possible combinations of priority connections which require a total outgoing line capacity of C_p or less. The transition rate matrix (or infinitesimal generator) Q of the Markov process is then $M \times M$, and grows rapidly as the number of different priority classes increases.

When the system is in state number i, the components of the state vector S_i are denoted by $m(i)$ and $p_n(i)$, $n = 1, 2, \ldots, K$. Then the entries Q_{ij} in the transition rate matrix are obtained as follows:

- If the state vectors S_i and S_j differ in more than one component, there are no direct transitions from i to j and $Q_{ij} = 0$.
- If S_i and S_j differ only in the first component (the SPP mode), then $Q_{ij} = \alpha_{m(i)}$, the SPP mode duration parameter for mode $m(i)$.
- If state j has one more priority connection of class n then $Q_{ij} = \lambda_n$, the average arrival rate of calls from class n.
- If state j has one less priority connection of class n, any of the $p_n(i)$ calls in progress may be completed and $Q_{ij} = p_n(i)\mu_n$.
- If $i = j$ (no change in state) then Q_{ii} is determined from

$$\sum_{j=1}^{K+1} Q_{ij} = 0 \ .$$

Note that the transition rate matrix Q does not involve the two SPP arrival rates η_0 and η_1. The stationary probability vector $\Pi = [\pi_1, \pi_2, \ldots, \pi_m]$ for the Markov process is computed from the transition rate matrix by solving simultaneously

$$\Pi Q = 0$$

and

$$\sum_{i=1}^{m} \pi_i = 1 \ .$$

For each state i, the capacity $C(i)$ available to non-priority traffic is known. For the queue to remain stable, the average non-priority throughput cannot exceed

$$\sum_{i=1}^{m} C(i)\pi_i \ .$$

This non-priority NB traffic queue can now be studied as a Quasi-Birth-and-Death (QBD) process whose state is specified by the number of queued messages and the state of the Markov environment. When the system is in state i, the average service rate for non-priority messages is

$$\gamma_i = \frac{C - C(i)}{L_{NP}} \ .$$

Given that no limit is assumed for the queue length, the QBD process has a semi-infinite generator whose entries are:

$$\tilde{Q} = \begin{bmatrix} Q - \Delta(\eta) & \Delta(\eta) & 0 & 0 & \dots \\ \Delta(\gamma) & Q - \Delta(\eta+\gamma) & \Delta(\eta) & 0 & \dots \\ 0 & \Delta(\gamma) & Q - \Delta(\eta+\gamma) & \Delta(\eta) & \dots \\ . & . & . & . & \\ . & . & . & . & \end{bmatrix}$$

where $\Delta(\eta)$ and $\Delta(\gamma)$ are $M \times M$ diagonal matrices with entries $\eta_{m(i)}$ and γ_i, $i = 1, 2, \dots, m$. The ith diagonal entries in $\Delta(\eta)$ and $\Delta(\gamma)$ represent the effective arrival and service rates for the non-priority queue when the system is in state i.

It is shown in [1] that the stationary joint probability of finding i messages in the non-priority queue and the Markov environment in state j, $j = 1, 2, \dots, m$ is a vector $X_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ which can be found by computing

$$X_i = \Pi(I - R)R^i, \quad i = 0, 1, \dots$$

where the $M \times M$ matrix R is the solution of

$$R^2\Delta(\gamma) + R[Q - \Delta(\gamma + \eta)] + \Delta(\eta) = 0 \ .$$

For computational purposes, R can be obtained by successive substitution starting with $R = 0$ in the rearranged formula

$$R = [R^2\Delta(\gamma) + \Delta(\eta)][\Delta(\eta + \gamma) - Q]^{-1} \ .$$

Once the vectors X_i are determined, the stationary probability of finding i messages in the non-priority queue is:

$$\sum_{j=1}^{m} x_{ij}$$

from which the average queueing delay can be found.

2.5 Results

This section gives the results for various strategies and different methods of analysis.

The appropriate measures of performance for the WB traffic is the blocking probability. For this traffic the aim of the control strategy should be to keep the blocking probability to an acceptable level. For NB calls, the average queueing delay is the best single measure of performance. For the combined traffic, these two measures have to be combined in some weighted fashion. Following [95], the overall merit of an algorithm will be computed from

$$\frac{1 - P_b}{\mu_1 E(t_s)}$$

where P_b is the WB blocking probability and t_s is the system delay for NB calls. The results presented in this section have been obtained for the following parameters:

$b_1 = 1$, $b_2 = 3$, $m_1 = 4$, $m_2 = 2$, $\mu_1 = 1$, $\lambda_2 = 1$, $\mu_2 = 1$ and $C = 10$.

The legend of each graph identifies the particular strategy used to obtain that graph. Other keys to interpreting the graphs are given in parentheses in front of the access strategies abbreviations:

- (s) indicates that the results are obtained from simulation.
- (MC) indicates that the results are obtained from the finite-state Markov chain analysis.
- (MG) indicates that the results are obtained from matrix-geometric methods.

Figures 2.5 to 2.9 show the performance results obtained for various strategies, using different methods of analysis. These figures show that the results of the three methods of analysis are almost identical. The expected advantages of statistical multiplexing are apparent in the comparison of Figures 2.5 and 2.8. For

the particular parameters used in these analyses, the minimum NB call system delay for the non-statistical multiplexing case is 1, which is equivalent to the predetermined fixed service rate μ_1. With statistical multiplexing, this quantity can be as low as about 0.1 for very light NB and WB loadings.

Between the two analytical methods used, the matrix-geometric method produces results considerably faster than the finite Markov chain method. The large number of states which are carried in the Markov chain analysis lead to much larger matrices than those needed, for example, to describe only the WB connection state for the matrix-geometric method. There are also clear advantages in the decomposition method of finding the invariant probabilities, and none of the problems of the sizes examined so far seem to require the slower but more accurate iterative techniques. Some typical computation times[2] for the MBP strategy are as shown in Table 2.1.

Procedure	*Remarks*	*Time(s)*
Simulation	SIMSCRIPT II.5	1200
Finite Markov	Iterative	27
Matrix-geometric	Decomposition	0.35
Matrix-geometric	Iterative	2.9

Table 2.1: Typical Computation Times

2.6 Summary

In this chapter we have considered some access control strategies for a non-statistical TDM multiplexer as well as for an ATM multiplexer. We have shown how different methods can be applied in the performance analysis of these systems. In section 2.3, several control strategies have been studied for an access node multiplexer that serves WB and NB traffic in a synchronous TDM environment. The strategies studied are MBNSD (movable boundary with no sorting of the channel allocations of the digital pipe), MB (movable boundary with sorting

[2]Scaled to a 25Mhz PC/AT with co-processor

of channel allocations of the digital pipe), and MBP (movable boundary with pre-emption).

These strategies have been analysed through different methods including simulation, an approximate Markov chain analysis based on iteration, and a decomposition method of matrix-geometric analysis. Results from various methods of analysis have been in agreement. These results indicate that the MBNSD strategy favours NB traffic the most, with the MB and MBP strategies after it. The order for favouring WB traffic is MBP strategy first, MB strategy second and MBNSD strategy third. A combined performance measure has indicated that MBNSD is a better strategy overall. Furthermore, MBNSD is also the easiest strategy to implement because it does not require the access node to have the capability of reassigning the channel allocations for NB calls in progress.

In section 2.4, the MBP strategy has been modified for the ATM environment and the performance of the system has been analysed using simulation methods, an approximate Markov chain analysis and an iterative method of matrix-geometric analysis. For the same set of traffic parameters, statistical multiplexing has shown an improvement in the performance of the NB traffic by a large factor as compared to synchronous TDM multiplexing (see Figures 2.5 and 2.8).

Among different methods of analysis, the decomposition method of matrix-geometric analysis has been found to be the fastest. None of the problems investigated have required the slower, but more accurate iterative techniques.

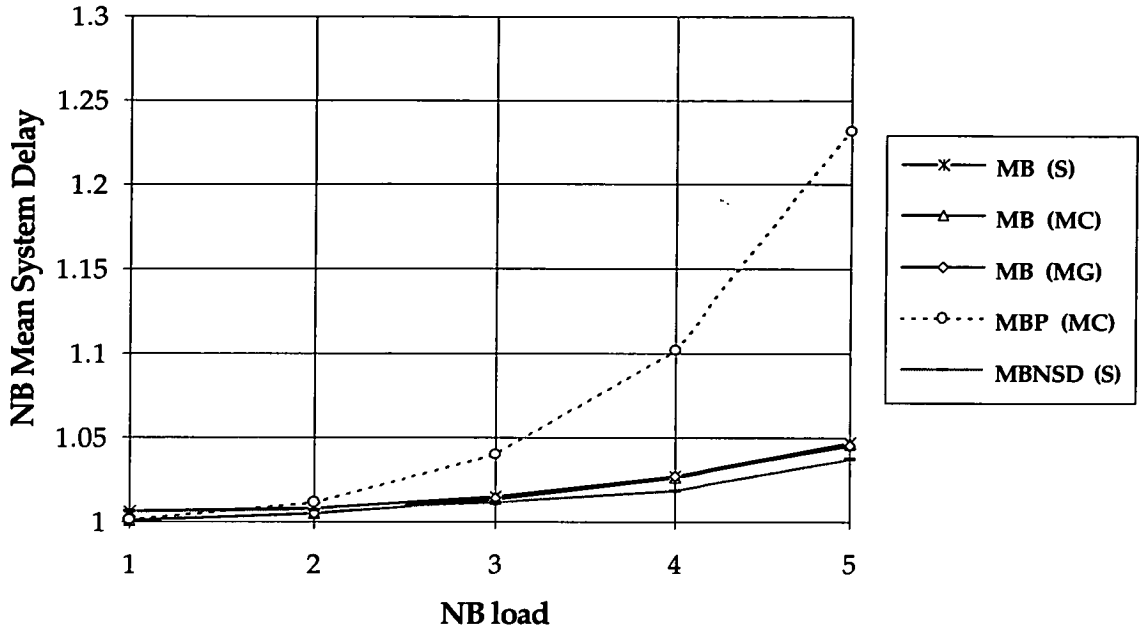


Figure 2.5: The system delay for NB traffic in TDM multiplexer

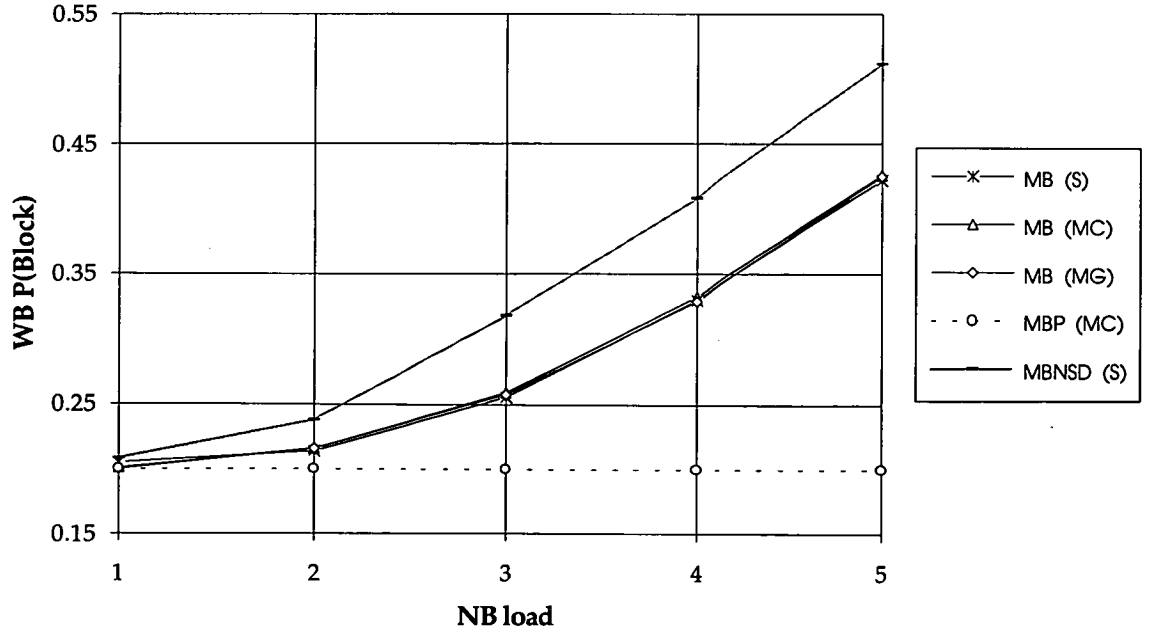


Figure 2.6: The blocking probability for WB traffic in TDM multiplexer

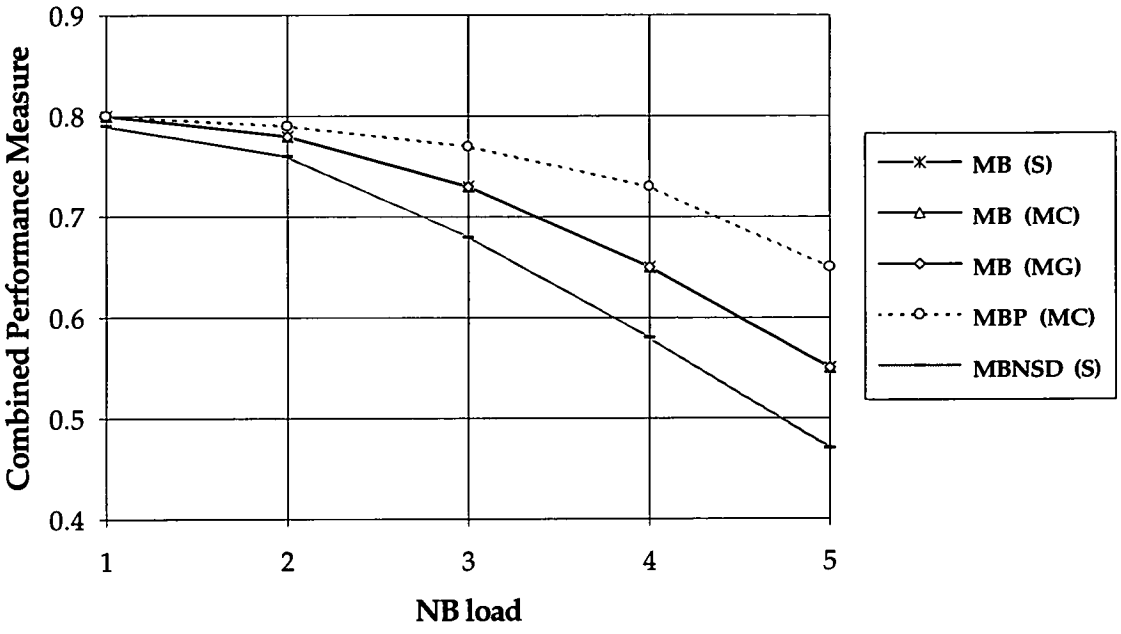


Figure 2.7: The combined performance measure in TDM multiplexer

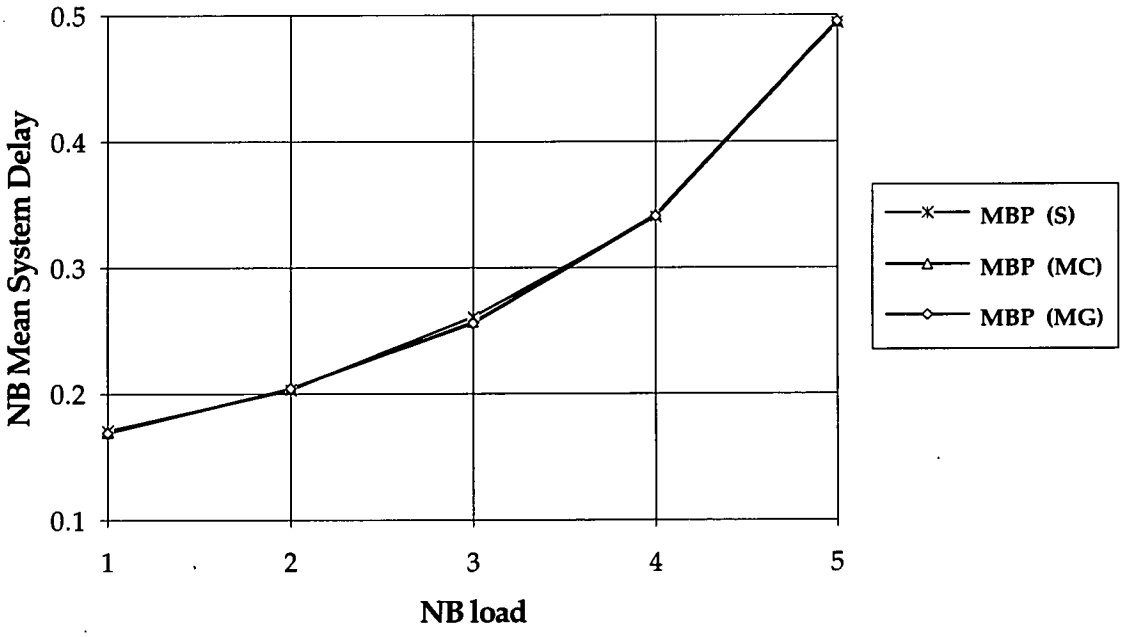


Figure 2.8: The system delay for NB traffic in ATM multiplexer

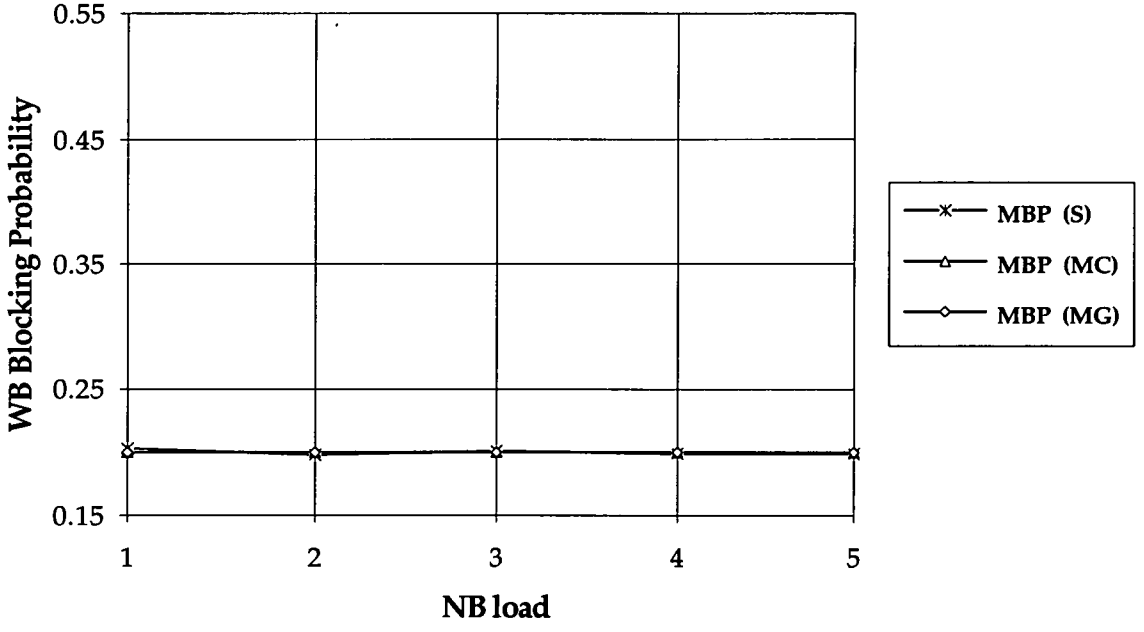


Figure 2.9: The blocking probability for WB traffic in ATM multiplexer

Chapter 3 has been removed for copyright or proprietary reasons.

Part of the work presented in this chapter has been published as: Habibi, D., Lewis, D.J.H., Nguyen, D. T. Access control in ATM networks carrying video, interactive images and data traffic, in, Australian Broadband Switching and Services Symposium, pages 165-174, Sydney, July 1991.

Chapter 4 has been removed for copyright or proprietary reasons.

The work presented in this chapter has been published as: Habibi, D., Lewis, D.J.H., Nguyen, D.T. Pieloor, J. Performance of a multiplexer in a B-ISDN network with STM and ATM traffic, in, Australian Broadband Switching and Services Symposium, pages 691-698, Melbourne, July 1992, and,
Habibi, D., Lewis, D.J.H., Nguyen, D.T. Pieloor, 1993. J. Analysis of an access node multiplexer in a system serving CBR & VBR traffic, Computer communications, 16(12)

Chapter 5

Traffic Models for Video Services

5.1 Introduction

Video services are considered to be dominant traffic elements in B-ISDN. In fact, video services will greatly influence the overall data rate requirements in ATM networks. Because of the very high bandwidth of video services, some steps must be taken to reduce the bit rate of the commonly used video services so that they can be handled more easily by the network. This would also make video services available to subscribers at a lower cost. There are several approaches that may be used to reduce the bandwidth requirements of video services:

- Apply data compression techniques to remove redundant information within a single video frame.
- Apply data compression techniques to remove interframe redundancies.
- Allow for some distortions, preferably those that are least detectable by the human visual system.

All the approaches outlined above emphasize the fact that bit rate reduction of video requires a good understanding of both the human visual system, and the nature of the particular video service. The acceptable quality for a video

service is defined based on its application. For example, in videotelephony, particularly for residential applications, the required resolution is low and the rate of change of picture is also very low. For high definition television (HDTV) however, near cinema quality is expected which is clearly a lot higher than the quality of videotelephony.

Table 5.1 shows five levels of quality for video services as defined by CCITT. This table also indicates the current state of digital encoding technology for such images.

Service Quality	*Description*	*Data Rate (Mbps)*
A	High definition television (HDTV)	92 - 200
B	Digital component-coding signal	30/45 - 145
C	Digitally-coded NTSC, PAL, SECAM for distribution	20 - 45
D	Reduced spatial resolution and movement portrayal	0.384 - 1.92
E	Highly-reduced spatial resolution and movement portrayal	0.064

Table 5.1: Bit Rates for Compressed Video Transmission

B-ISDN provides a flexible transport environment for data, voice and video. The dynamic allocation of bandwidth in B-ISDN makes it possible to replace the conventional CBR video transmission with VBR video transmission which has the advantages of stable picture quality and better efficiency through statistical multiplexing. With the introduction of VBR video, a new area of research has evolved in relation to the mathematical modelling of video traffic [104, 105, 106].

One issue in the performance analysis of video traffic is the way in which video traffic is transmitted in each frame. For a case where several video sources are multiplexed [107], it has been shown that the worst multiplexer performance occurs when at the beginning of each frame, the source transmits at peak rate until the information for that frame is exhausted. This produces an inherent correla-

tion between video sources because they all have the same periodic on/off sample path. Hence, synchronisation between video sources becomes an uncontrollable factor in the performance of the multiplexer. The best case for multiplexer performance is reported to be when the information within a frame is distributed in time over that frame [107]. This reduces the structural correlation between sources.

Video transmission generates traffic exhibiting both short term and long term correlated cell arrivals. The fast-decaying short term correlation corresponds to uniform activity levels with a time constant in the order of a few hundred milliseconds. The slow-decaying long term correlation corresponds to sudden changes in gross activity level of the scene (in other words sudden scene changes), and its time constant is in the order of a few seconds [108].

5.2 Models Considering Only Short Term Correlations

In this section we look at those models which only take into account the short term correlations, i.e. for video sources without scene changes. An example of such video source is videotelephony where the scene is typically a person's face. Two models are examined in this section. The first model is the 'continuous-state autoregressive Markov model' [104, 105]. This model characterises time domain behaviour of video information by autocorrelation and approximates a single video source by the autoregressive (AR) process [105]. The second model is the 'discrete-state, continuous-time Markov process' [104]. This model approximates the source rate by a continuous time process with discrete jumps at random (Poisson) time instants.

5.2.1 Model A: Continuous-State Autoregressive Markov Model

In this model a single video source is approximated by the autoregressive (AR) process [104]. An autoregressive Markov process, $\lambda(n)$, is generated by the recur-

sive relation [105]:

$$\lambda(n) = \sum_{m=1}^{M} a_m \lambda(n-m) + bw(n) \tag{5.1}$$

where $\lambda(n)$ represents the source bit rate during the n^{th} frame, $w(n)$ is a sequence of independent Gaussian random variables, M is the order of the model, and a_m $(m = 1, 2, \cdots, M)$ and b are constant coefficients. If we assume a first order model then[104]:

$$\lambda(n) = a\lambda(n-1) + bw(n) \ . \tag{5.2}$$

It is assumed that $w(n)$ has mean η and variance 1. It is also assumed that $|a| < 1$. Thus, the process is able to achieve steady state with large n. The steady-state average $E(\lambda)$ and discrete autocovariance $C(n)$ are given by [90]:

$$E(\lambda) = \frac{b}{1-a}\eta \tag{5.3}$$

$$C(n) = \frac{b^2}{1-a^2}a^n \quad n \geq 0 \ . \tag{5.4}$$

The values of the coefficients a and b can be determined by matching the steady-state average $E(\lambda)$ and the discrete covariance $C(n)$ of the AR process with the measured data. It is shown in [105] that with a first order autoregressive process, the bit rate probability density function for a video source is accurately modelled and that the autocorrelation curve of the video in the short term (up to 0.5 s) is also accurately modelled. A higher order model is required to approximate high autocorrelation over 1 second [105]. It should be noted that this video model is easy to simulate but queueing analysis of it becomes very complex. Even a continuous flow approximation of the queueing process with the input as given in equation (5.2) leads to a two-dimensional diffusion partial differential equation, with reflecting barriers at zero for both the input rate and the queue size [104].

5.2.2 Model B: Discrete-State, Continuous-Time Markov Process

In this model the bit rate is quantised into finite discrete levels [104]. Transitions between levels are assumed to occur at exponentially distributed times that may depend on the current level. Hence the source rate is approximated by a

continuous-time process $\bar{\lambda}(t)$ with discrete jumps at random (Poisson) times [104]. An example of this is shown in Figure 5.1 [104]. Clearly, increasing the sampling rate and decreasing the quantisation step, A, will improve the accuracy of the approximation. It should be noted however that these parameters are related. For example, to take advantage of a smaller quantisation step, the sampling rate needs to be increased.

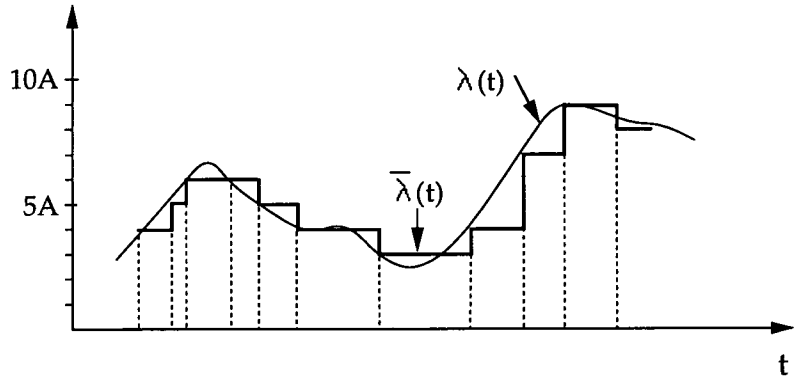


Figure 5.1: Poisson sampling and quantisation of the source rate

Note that here, the measured rate is treated as a continuous-time, continuous-state process $\lambda(t)$, because, compared to the time scale, the frame period is very small. This model can be used to describe a single source as well as an aggregate of N independent sources each with rate $\lambda(t)$, mean $E(\lambda)$ and autocovariance $C(\tau) \simeq C(0)e^{-a\tau}$ at steady state. In this case the parameters of the model must be tuned to the aggregate instantaneous rate $\lambda_N(t)$. The steady-state mean and covariance of $\lambda_N(t)$ are given by:

$$E(\lambda_N) = NE(\lambda) \tag{5.5}$$

$$C_N(\tau) = NC(0)e^{-a\tau} \ . \tag{5.6}$$

In [104] it is suggested that a birth-death Markov model will accurately describe the aggregate source bit rate. It is further expected that the tendency of the bit rate toward higher levels decreases at high levels, and, the tendency of the bit rate toward lower levels increases at high levels. The resulting stationary distribution of the state has a bell shape. Figure 5.2 [104] shows the transition diagram of a simple birth-death process that exhibits this behaviour and which

has an exponential autocovariance. The diagram shows $M+1$ possible levels, and uniform quantisation of A bits/pixel is assumed. The exponential transition

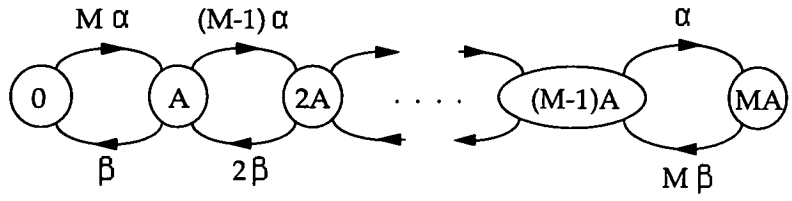


Figure 5.2: State transition diagram - Model B

rates $r_{i,j}$ from state iA to state jA are given by:

$$\begin{aligned} r_{i,i+1} &= (M-i)\alpha \quad i < M \\ r_{i,i-1} &= i\beta \quad i > 0 \\ r_{i,j} &= 0 \quad |i-j| > 1 \ . \end{aligned}$$

It is shown in [109] that $\lambda_N(t)$ has a binomial distribution with mean $E(\lambda_N)$, variance $\bar{C}_N(0)$ and exponential autocovariance $\bar{C}_N(\tau)$ at steady state, with

$$\begin{aligned} P\{\lambda_N(t) = kA\} &= \binom{M}{k} (\frac{\alpha}{\alpha+\beta})^k (1-\frac{\alpha}{\alpha+\beta})^{M-k} \\ &= MA\frac{\alpha}{\alpha+\beta} \end{aligned} \tag{5.7}$$

$$\bar{C}_N(0) = MA^2\frac{\alpha}{\alpha+\beta}(1-\frac{\alpha}{\alpha+\beta}) \tag{5.8}$$

$$\bar{C}_N(\tau) = \bar{C}_N(0)e^{-(\alpha+\beta)\tau} \tag{5.9}$$

where α and β are the transition rates. The parameters M, A, α and β are obtained by matching the above equations with the measured data. The process in Figure 5.2 can be thought of as a superposition of M independent identical minisources. Each minisource alternates between transmitting 0 bits/pixel and A bits/pixel as shown in Figure 5.3 [104].

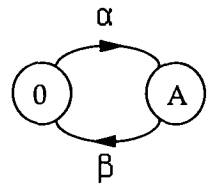


Figure 5.3: Minisource model

5.3 Models Considering Long Term Correlations

In this section we look at those models which also take into account the long term correlations as well as short term correlation, i.e. for video sources with scene changes. Five models are examined in this section. The first model is an extension of Model B detailed earlier [108]. The second model [110], also derived from Model B, has three motion activity classes. Three different models are used for each class and the transition probabilities from one class to the other are measured from the actual video data. The third model which will be only briefly outlined approximates a video source by a discrete-state, continuous-time Markov process with batch arrivals [111, 112]. The fourth model [113] is based on the Transform-Expand-Sample methodology and claims several advantages over the first order autoregressive models [113]. The last model is a histogram based model [114] and considers a multiplexer that serves video traffic. This model finds the buffer occupancy distributions for all sources and uses them to approximate the buffer occupancy of the multiplexer.

5.3.1 Model C: An Extension of Model B for Video Sources with Scene Changes

This extended model [108] which includes both short term and long term correlations, involves a correlated Markov process model with a state transition rate diagram as shown in Figure 5.4 [108].

The source is represented as one which changes among different fixed rate levels. Each state has been labelled by the data rate of the prebuffer corresponding to

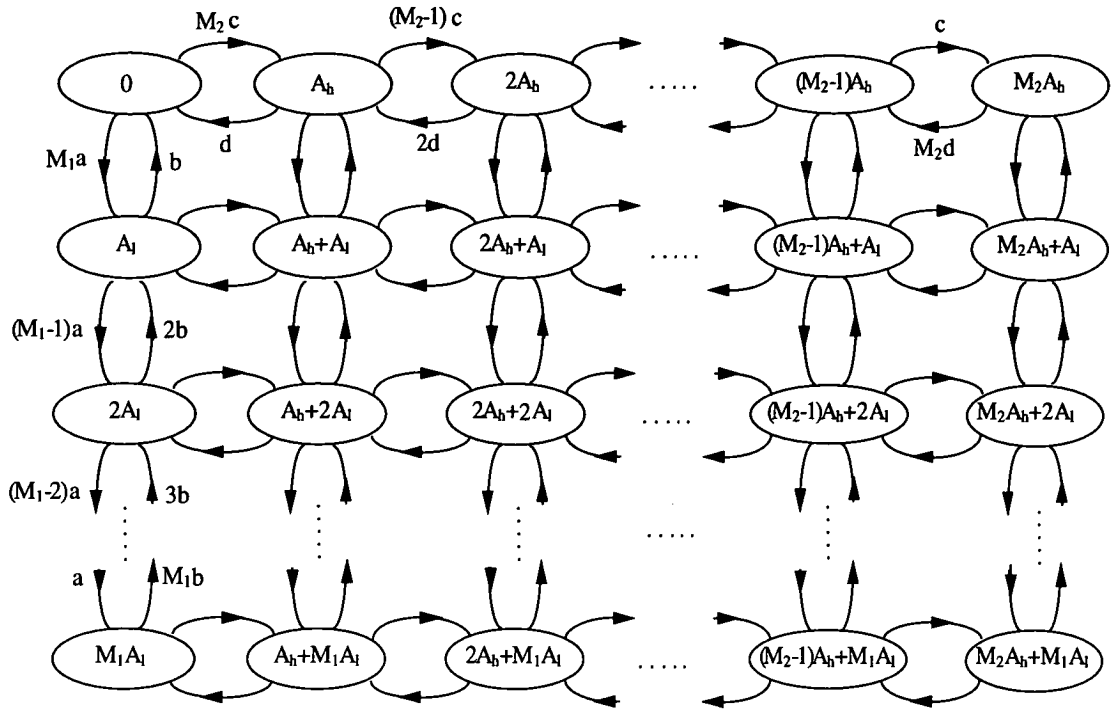


Figure 5.4: State transition diagram - Model C

that state. All data rates are integer combinations of two basic levels: A_h, the high rate, and A_l, the low rate. The high data rate A_h represents a sudden scene change and the low data rate A_l represents a uniform activity level (i.e. small fluctuations in the bit rate). There can be a maximum of $(M_1 + 1)$ low rate levels and a maximum of $(M_2 + 1)$ high rate levels for an aggregate process of N video sources. Thus, there are $(M_1 + 1)(M_2 + 1)$ different levels among which the aggregate process can transmit, where for an arbitrary value of M, $M_1 = NM$ and $M_2 = N$. If scene changes do not exist then we can delete all the states which contain a high rate A_h in which case Figure 5.4 reduces to Figure 5.2 which was used previously for Model B.

The state transition probabilities between uniform activity level and high activity level can be determined from the actual measured data, i.e. c and d are determined by equating the fraction of time spent in the high activity level $(c/(c+d))$ and the average time spent in the high activity level $(1/d)$ with the actual measured data. To determine the transition probabilities within the uniform activity

level (a and b), the high data rate A_h, and the low data rate A_l, the first and second order statistics are matched with the actual measured data.

In the same way that Model B could be decomposed as a superposition of M independent identical minisources, Model C can also be decomposed as a superposition of M_1 independent identical minisources of Figure 5.5(a) and M_2 independent identical minisources of Figure 5.5(b) [108]. Then, in order to specify the state of the system, it is sufficient to know the number of each type of minisource in the ON state.

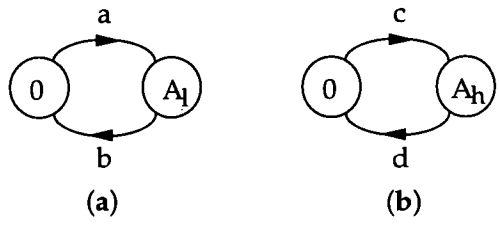


Figure 5.5: Minisource models

5.3.2 Model D: Multi-Level Continuous-State Autoregressive Markov Model

This model considers three motion activity classes for modelling of motion classified VBR video codecs [110]. The three motion activity levels are low, medium and high motion. Each of the motion activity levels is modelled by a first order autoregressive model:

$$\lambda_i(n) = a_i\lambda_i(n-1) + G_i(n) \qquad (i = 1, 2, 3) \tag{5.10}$$

where $i = 1$ refers to low motion, $i = 2$ refers to medium motion and $i = 3$ refers to high motion activity levels. $\lambda_i(n)$ denotes the number of bits generated from the video codec at the n^{th} frame of class i, a_i is a constant for class i and $G_i(n)$ is a Guassian random variable with mean η_i and variance σ_i^2. The three autoregressive processes are used to cater for various activity levels in the video. The duration of each class in terms of the number of consecutive frames for which

the video traffic falls within that class has a geometric distribution. The process that describes the transition between different motion activity classes is a discrete time Markov process. The density function of the duration of each class is given by the following geometric distribution:

$$F_i(k) = \frac{\theta_i}{1-\theta_i}(1-\theta_i)^k \tag{5.11}$$

where k is a random variable that represents the duration of a class and has a mean of $1/\theta_i$. Taking π_{ij} to be the transition probability from the current class, i, to the next class, j, then the matrix that gives the transition probabilities of moving from class i in the current frame to class j in the next frame is given by $\mathbf{P} = [p_{ij}]$ and is represented as follows:

$$\mathbf{P} = \begin{bmatrix} 1-\theta_1 & \theta_1\pi_{12} & \theta_1(1-\pi_{12}) \\ \theta_2(1-\pi_{23}) & 1-\theta_2 & \theta_2\pi_{23} \\ \theta_3\pi_{31} & \theta_3(1-\pi_{31}) & 1-\theta_3 \end{bmatrix}. \tag{5.12}$$

This three class video model is completely specified by a_i, η_i, σ_i^2, θ_i and π_{ij}.

5.3.3 Model E: Discrete-State, Continuous-Time Markov Process with Batch Arrivals

In this model [111, 112], uniform activity level is described by a discrete-state, continuous-time Markov process similar to Model B. Batch arrivals are used to describe sudden scene changes (e.g. scene changes in broadcast TV). The batch size is assumed to be constant and the interarrival time between batches is assumed to be exponentially distributed.

5.3.4 Model F: Transform-Expand-Sample Based Model

This model is proposed in [113] and is based on the Transform-Expand-Sample (TES) methodology proposed in [115] and [116]. TES is a class of methods for generating correlated variates through autoregressive modulo-1 schemes and have several advantages [113]. Firstly, the TES methods are non-parametric which means that they can generate any marginal distribution or an arbitrarily close approximation. This makes TES suitable for modelling empirical data because

it can generate a marginal distribution which exactly matches an empirical histogram. Secondly, the associated autocorrelation function can easily be computed hence facilitating a search for a fit. Thirdly, TES methods are very fast. Uniform TES sequences have slightly higher computational complexity than the underlying pseudo-random number generator. Finally, TES sequence behaviour is very versatile, ranging from driftless random walks to cyclic random walks. The video model proposed in [113] is quite detailed and cannot be easily summarised. Full details of the model may be found in the reference. In [113] it is claimed that this model has some advantages over the first order autoregressive models.

5.3.5 Model G: A Histogram Based Model

This model is based on a source bit rate histogram [114]. A multiplexer is considered that serves video traffic. Because of the fixed ATM cell size, the ATM cell generation on a frame by frame basis is approximated by an $M/D/1/N$ queue. The buffer occupancy of this system can be computed to an arbitrary degree of precision using an $M/E_k/1/N$ approximation by increasing the value of k [117]. For a case where a single source enters the multiplexer, the arrival of ATM cells in any frame is assumed to be a Poisson process with a rate of λ. Over the whole sequence, λ is a random variable with distribution $f_\lambda(x)$. It is assumed that because of the large number of cells in each frame, the system would reach steady state very quickly, compared to the duration of the frame. Hence it is assumed that by solving the $M/D/1/N$ problem as a function of λ and then conditioning over the range of λ, the statistics of the system may be found.

The buffer occupancy distribution for the given source would be:

$$P(n) = \sum_{i=1}^{M} P(n|\lambda_i)P(\lambda_i) \tag{5.13}$$

where M is the number of intervals in the histogram, $P(n|\lambda_i)$ is the buffer occupancy distribution given the arrival rate is λ_i, and $P(\lambda_i)$ is the histogram approximation of $f_\lambda(x)$. When several such sources are multiplexed, the buffer occupancy approximation may be calculated as the weighted sum of the individual buffer occupancy distributions. This has been shown in Figure 5.6 [114].

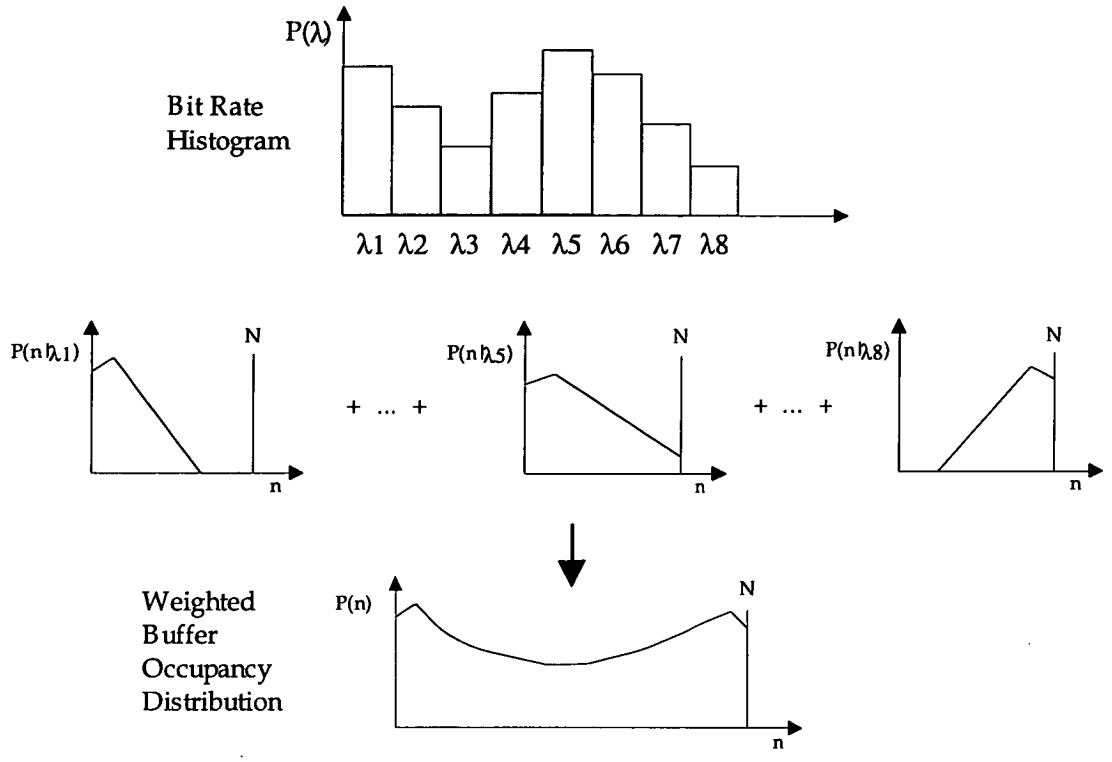


Figure 5.6: A weighted buffer occupancy distribution calculated using the histogram model

5.4 Summary

In this chapter we have discussed the importance of video traffic modelling. Several important issues in video traffic modelling have been outlined. A literature survey has been provided for those models that only take into account short term correlation in the traffic generated by video, as well as other models, that also take into account the long term correlation, i.e. videos with a lot of scene changes.

Chapter 6

Performance of Hidden Markov Models for VBR Video Traffic

6.1 Introduction

As stated in the last chapter, video services are considered to be one of the dominant traffic elements in B-ISDN. The wide range of future video services and their bandwidth relative to non-video services suggest that video traffic will effectively control the overall performance and data rate requirements in ATM networks.

Depending on application, video traffic exhibits both short term and long term correlations in the pattern of cell arrivals. An good video model must take into account such correlations and yet be simple enough to be mathematically tractable. Of various models for video services, those which embody an underlying Markovian mechanism are likely to lead to an analytical solution. In the last chapter several models which have been proposed for various video services were outlined. In this chapter we further investigate models which are suitable for VBR video. The models considered in this chapter are based on hidden Markov models. The applicability of these models to modelling VBR video traffic is investigated and some performance results are presented to show the accuracy of these models for queueing purposes. The work presented in this chapter has been published by Habibi in [118].

6.2 Hidden Markov Models

We begin brief description of Markov models and from there proceed to hidden Markov models. A Markov process is a stochastic process whose dynamic behaviour is such that the probability distribution for its future developments depends only on the present state and not on how the process arrived at that state [93]. A Markov process could be called observable if the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable) event [119].

An Observable Markov Process Example: As an example of an observable Markov chain let us consider the average day time temperature of Hobart during summer (all numbers given in this example are purely fictitious). Let us define three temperature states as follows:

State	*State Index*	*Temperature Range* (°C)
Cool	1	< 15
Mild	2	15-22
Hot	3	> 22

We assume that Hobart's weather has the following state transition probabilities matrix:

$$P = \{p_{ij}\} = \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.7 & 0.2 \end{bmatrix}$$

The state transition diagram for this example is shown in Figure 6.1. In this example the Markov model is observable because each state corresponds to an observable event.

An observable Markov model is too restrictive to be applicable to many problems of interest. Consider a case where the observation is a probabilistic function of the state, i.e. the resulting model which is called the hidden Markov model is a doubly embedded stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of

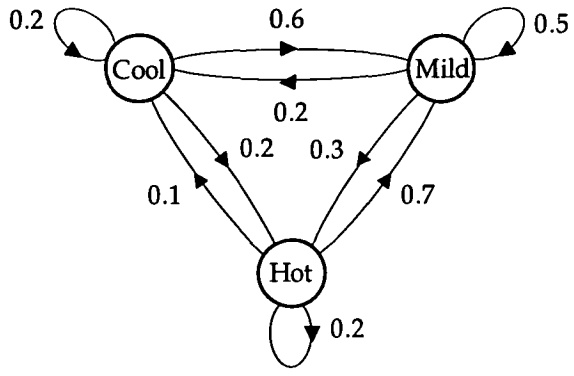


Figure 6.1: A Markov chain representing the transitions in Hobart's climate

stochastic process that produce the sequence of observations.

The theory of hidden Markov models was first published in late 60's and early 70's in a series of papers by Baum et al. [120, 121, 122]. Hidden Markov models attracted a lot of interest in the area of speech recognition in the 70's and the 80's [123, 124, 125, 126, 127]. A good tutorial on hidden Markov models is given in [119] where it is also shown how it can be applied to speech recognition. Let us illustrate the concept of hidden Markov models by providing the following two examples:

Hidden Markov Model Example (I): Consider the following scenario. A person is sitting at a table. There are three hats on the table each containing a mixture of coloured dices. The colours consist of white, yellow, green, orange, red, and violet. Each hat contains a different distribution of colours. Initially, the person selects one of the hats at random. He has to make N trials on the current hat. Without looking inside the hat the person picks out a dice from the hat and announces his observation, that is the colour of the dice. He puts the dice back in the hat and makes another trial. This procedure is repeated until N trials are completed. A new hat is then selected based on a random selection process (defined by a transition matrix) which is a function of the current hat. N trials are made on the new hat and another hat is selected. Let us define the state of the process as the hat on which the trials are made. The observation of the process is the sequence of colours announced by the person and can be mod-

elled by knowing state transition probabilities (probabilities for determining the sequence of hats used for the experiment) and the probability density function of colours within each hat. This example is illustrated by Figure 6.2.

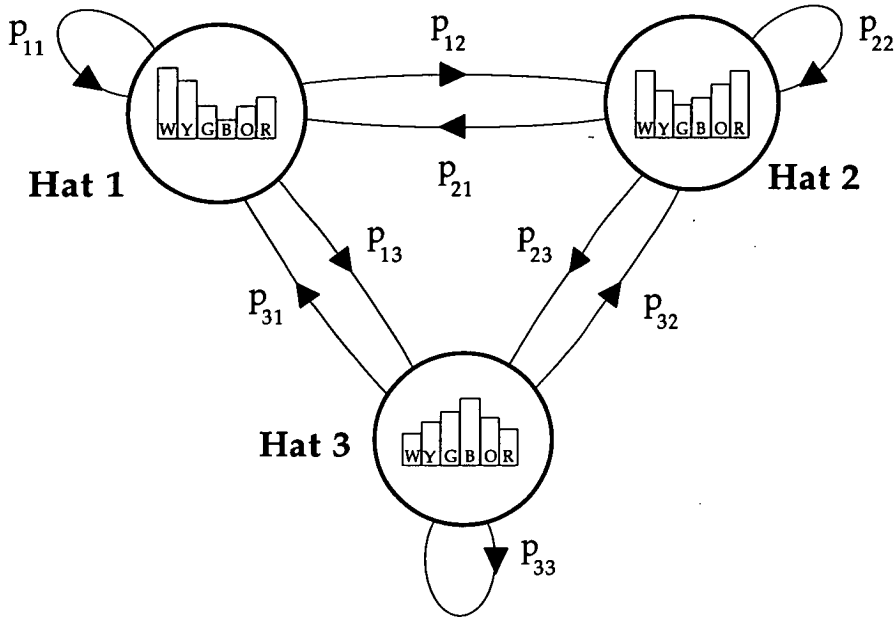


Figure 6.2: An illustration of the hidden Markov model of Example (I)

Hidden Markov Model Example (II): This is a similar scenario to Example (I). The selections among hats are made with the same procedure as the last example. The difference is the contents of the hats and the actual trial procedure. Each hat now contains 6 dices, one of each colour. Some of the dices within each hat are biased. A biased dice is one that when tossed, the probabilities of observing numbers 1 to 6 are not equivalent. The bias of the dices of the same colour in different hats may be different. For example the red dice in hat 1 may have two '4' sides and no '6' side but the red dice in hat 2 may have three '6' sides and no '1' and '2' sides. The bias of each dice is known for each hat. Each of the six colours is associated with a unique number between 1 to 6. In this instance the following association is defined between colours and numbers:

Number	1	2	3	4	5	6
Associated colour	white	yellow	green	blue	orange	red

As in the last example, a person has to make N trials on each hat before proceeding to the next hat. For the first trial, one of the 6 dices in the hat is selected at random and its colour is announced. That dice is tossed once and depending on the outcome (numbers 1 to 6) the next colour to be tossed is determined. The first dice is then put back into the hat, and the second dice (which has been determined from tossing the first) is taken out of the hat, its colour is announced and then it is tossed. This process continues for N trials in the current hat before proceeding to select the next hat. The state of the system is defined as the hat which is being used for the trials and the sequence of observations is the colours announced by the person.

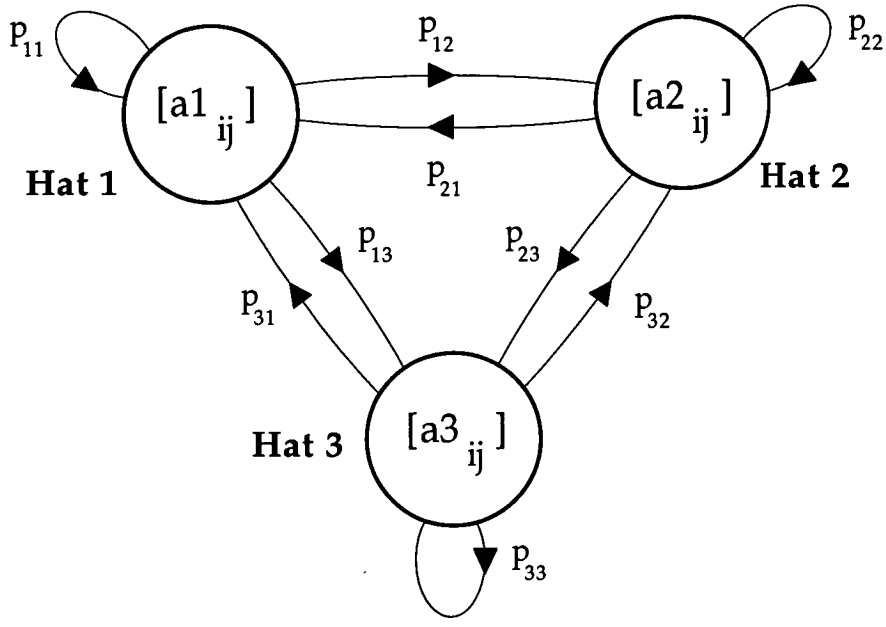


Figure 6.3: An illustration of the hidden Markov model of Example (II) (i and j range from 1 to 6)

In order to model the sequence of observations of this example we need to know the transition probabilities between the three hats, as well as the bias of the dices in each hat. When the biases of all 6 colour dices in a hat are known, a 6×6 colour transition probabilities matrix can be defined for that hat. The state (hat) transition probabilities matrix, plus the three 6×6 colour transition probabilities

matrices for the three hats, denoted by $[a1_{ij}]$, $[a2_{ij}]$, and $[a3_{ij}]$, completely specify the model. This example is illustrated in Figure 6.3.

6.3 HMM: A Hidden Markov Model for modelling VBR video

In a recent work, McLaren [128] uses a hidden Markov model for modelling the output cell stream of a video codec for *still images*. In this section we will outline a hidden Markov model that can be used for modelling the ATM cell stream generated from VBR motion video. The HMM implemented in this section has a lot of similarities to Example (II) of section 6.2. It consists of an *'outer'* (hidden) Markov chain whose states (in this context referred to as modes) are stochastic processes that produce the observation sequence. The observation in this context refers to the generation of ATM cells.

6.3.1 Traffic Generation and Modelling Procedure

Blocking and Coding of the Video

Video compression and coding is not within the scope of this thesis thus for the purposes of illustration we have considered a video coding scheme described in [97]. A summary of the coding scheme is shown in Figure 6.4.

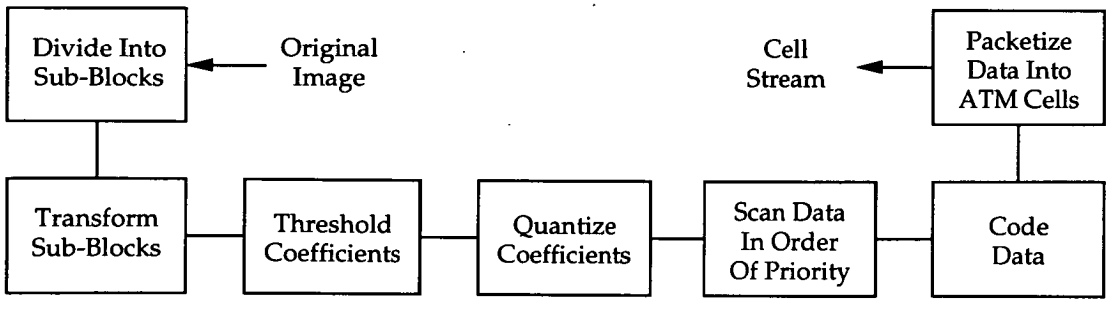


Figure 6.4: A VBR Layered Coding Scheme

Each original image file is divided into 16 × 16 pixels blocks. Then using a discrete cosine transform (DCT), the image subblocks are converted to blocks

of transform coefficients. These blocks of coefficients are then psychovisually thresholded and quantised. The transform coefficients are then scanned in order of increasing frequency and Huffman coded. The resulting bit-stream is packed into ATM cells in a non-priority manner.

Grouping of Blocks into Subframes

The image blocks resulting from dividing each image frame into 16 × 16 pixel partitions are grouped into subframes of size N, i.e. each subframe consists of N blocks of 16 × 16 pixels each. The value of N is taken to be a parameter of the model and must be selected such that a single video frame consists of an integer number of subframes. The concepts of block and subframe are illustrated in Figure 6.5.

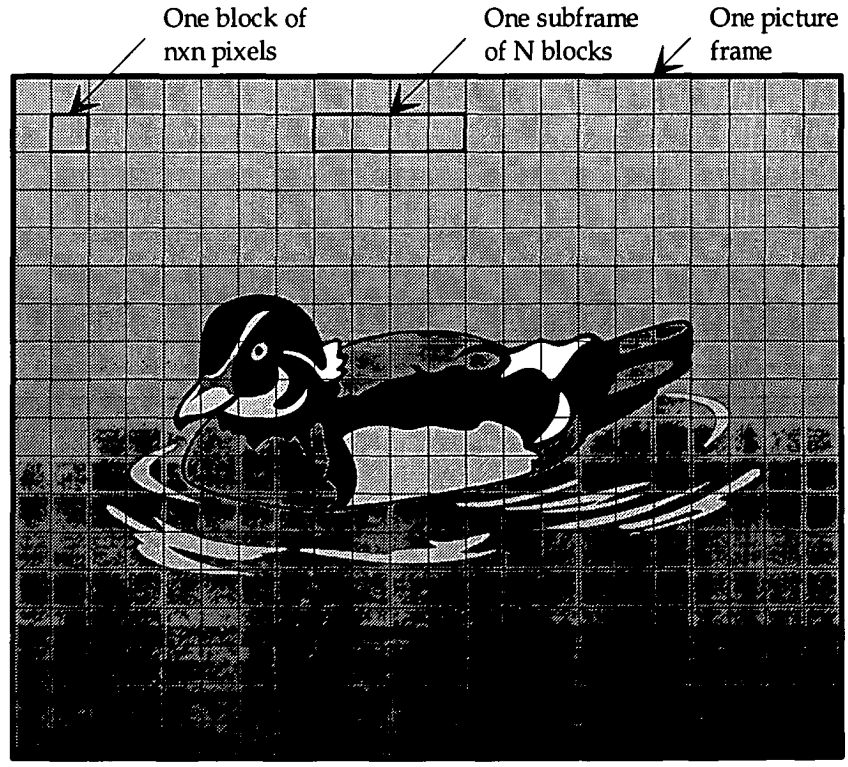


Figure 6.5: The relationship between blocks and subframes in a picture frame

Mode Definition, Modelling and Transitions

To cope with the wide variation in local complexity of the picture being scanned, and the corresponding variations in the effective number of cells per subframe produced, a number of *modes* are identified. Each mode corresponds roughly to a particular average level of fine detail in a local area of the picture. A mode in this context is the state of the *outer* (hidden) Markov process. The criteria for defining a mode is the number of ATM cells generated by a subframe. Depending on the number of ATM cells generated in a subframe, a mode is associated with that subframe. For specifying various modes of the model, a function (look-up table) is required that maps the number of cells generated from a subframe to a particular mode.

In the HMM approach, probabilistic transitions are assumed to take place between these modes as the scan proceeds, thus generating a Markov process described by a probability transition matrix. Therefore, if there are m modes specified for the video traffic, a $m \times m$ matrix, P, is necessary for the transition probabilities between different modes. Such a matrix will be of the form:

$$P = \{p_{ij}\} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1m} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2m} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & p_{m3} & \cdots & p_{mm} \end{bmatrix} \tag{6.1}$$

where p_{ij} is the probability of moving to mode j in the next subframe if the current subframe is in mode i. This matrix shall be called the intermode transition probabilities matrix.

Within each mode, the number of ATM cells produced from each of the blocks belonging to that subframe is also random, with the parameters of this distribution varying from mode to mode to suit the local complexity of the picture. Thus, if the maximum number of cells generated from a block is c, then a $(c+1) \times (c+1)$ matrix is necessary to specify the cell generation transition probabilities on a block by block basis for a particular mode. Therefore there should be m such

matrices denoted by $A_0, A_1, A_2, \cdots A_{m-1}$, for the m distinct modes where

$$A_k = \{ak_{ij}\} = \begin{bmatrix} ak_{00} & ak_{01} & \cdots & ak_{0c} \\ ak_{10} & ak_{11} & \cdots & ak_{1c} \\ \vdots & \vdots & \ddots & \vdots \\ ak_{c0} & ak_{c1} & \cdots & ak_{cc} \end{bmatrix}. \tag{6.2}$$

The entry ak_{ij} in the above matrix is, given the mode k, the probability of generating j ATM cells in the next block of the subframe if the current block has generated i ATM cells. These matrices are referred to as the *intramode cell generation transition probabilities matrices*, or *intramode transition matrices* for short.

6.3.2 Model Implementation and Verification

In our investigations of the HMM model, rather than finding the parameters of the HMM for single frames, we attempt to fit the model to a large sequence of motion video frames. We use the *Salesman* sequence of consecutive images, obtained from ftp site 128.113.14.50 under directory /pub/image/sequence/salesman/gray, to calculate the intermode transition matrix and the intramode transition matrices for the whole sequence. The dimensions of the Salesman frames were 288×360 pixels. All frames were trimmed to bring them to the nearest standard frame size of 288×352 pixels. Dividing each of these frames into blocks of 16×16 pixels results in 396 blocks per frame. Assuming a frame rate of 25 frames per second, the bit rates of the Salesman sequence before and after being subjected to the coding scheme is shown in Table 6.1. Note that the bit rate after coding includes the ATM cell headers.

Frame Rate	*Bit Rate Before Coding*	*Bit Rate After Coding*
25	20.28 Mbps	1.94 Mbps

Table 6.1: Bit rates of the Salesman sequence before and after coding

The maximum number of cells generated from any one block was 1. Therefore,

intramode matrices of dimensions 2×2 are sufficient to represent different modes. Although N is taken to be a variable, it is desirable to choose N such that one picture frame consists of a whole number of subframes. After choosing the value of N, the actual video data is processed to generate the probability density function of the number of cells generated from a subframe. These results are shown in Table 6.2 for subframe sizes of 2, 4, 6, 8, and 11 blocks.

	N=11	*N=8*	*N=6*	*N=4*	*N=2*
Cells per subframe	*Prob.*	*Prob.*	*Prob.*	*Prob.*	*Prob.*
0	0.00563	0.00715	0.01262	0.04472	0.18383
1	0.02214	0.03689	0.07046	0.23081	0.70974
2	0.00167	0.05921	0.25912	0.56284	0.10642
3	0.04423	0.26276	0.46392	0.15779	0
4	0.20650	0.44835	0.18218	0.00382	
5	0.32776	0.16161	0.01154	0	
6	0.29720	0.02272	0.00013		
7	0.09075	0.00126	0		
8	0.00408	0			
9	0				

Table 6.2: Pdf of the number of cells generated per subframe of the actual video

From the probability density function of the number of cells generated in a subframe, the number of modes, m, is chosen according to the following criteria:

- Where possible, the number of modes is chosen to be the same as the size of subframe (i.e. take $m = N$). If necessary, in Table 6.2, combine rows which correspond to very small probabilities into a single mode.
- Rows that correspond to high probabilities of occurrence are not combined.

For the size of the subframe, values of 2, 4, 6, 8, and 11 blocks were examined. Table 6.3 shows how modes have been defined for various sizes of subframe. Based

on these definitions, the parameters of HMM are calculated by fitting the model to the actual data.

	N=11		*N=8*		*N=6*	
Cells per subframe	*Mode index*	*Prob.*	*Mode index*	*Prob.*	*Mode index*	*Prob.*
0	0	0.00563	0	0.00715	0	0.01262
1	1	0.02214	1	0.03689	1	0.07046
2	2	0.00167	2	0.05921	2	0.25912
3	3	0.04423	3	0.26276	3	0.46392
4	4	0.20650	4	0.44835	4	0.18218
5	5	0.32776	5	0.16161	5	0.01154
6	6	0.29720	6	0.02272	5	0.00013
7	7	0.09075	7	0.00126		0
8	8	0.00408		0		
9	9	0				
10	10	0				

	N=4		*N=2*	
Cells per subframe	*Mode Index*	*Prob.*	*Mode Index*	*Prob.*
0	0	0.04472	0	0.18383
1	1	0.23081	1	0.70974
2	2	0.56284	2	0.10642
3	3	0.15779		0
4	3	0.00382		
5		0		

Table 6.3: HMM mode assignment for various values of N

For example, the matrices which completely specify the Hidden Markov Model

(HMM) of the Salesman sequence for N=4 are as follows:

$$P = \begin{bmatrix} 0.158874 & 0.340875 & 0.174460 & 0.325792 \\ 0.096101 & 0.154386 & 0.510331 & 0.239181 \\ 0.023503 & 0.234671 & 0.615437 & 0.126389 \\ 0.013641 & 0.296075 & 0.562222 & 0.128062 \end{bmatrix}$$

$$A_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 0.598729 & 0.401271 \\ 0.990163 & 0.009837 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0.194018 & 0.805982 \\ 0.823360 & 0.176640 \end{bmatrix} \quad A_3 = \begin{bmatrix} 0.011875 & 0.988125 \\ 0.413213 & 0.586787 \end{bmatrix} .$$

After the parameters of the model are determined for various subframe sizes, these models are used to generate traffic, and the statistics of model traffics are compared with those of the actual video traffic. The data of the actual video and those of the models are processed to generate results for

- the probability density function of the number of cells generated per subframe,
- the mean queue population if the traffic is fed to a dedicated server,
- the standard deviation of queue population,
- the normalised autocorrelation of cells generated from the actual video data and from various models.

Figures 6.6 to 6.10 show the probability density functions of the number of cells generated in a subframe from the actual video traffic and also from the HMM for different sizes of subframe. They indicate that the actual data and the HMM data give similar results for the probability of generating a certain number of cells per subframe.

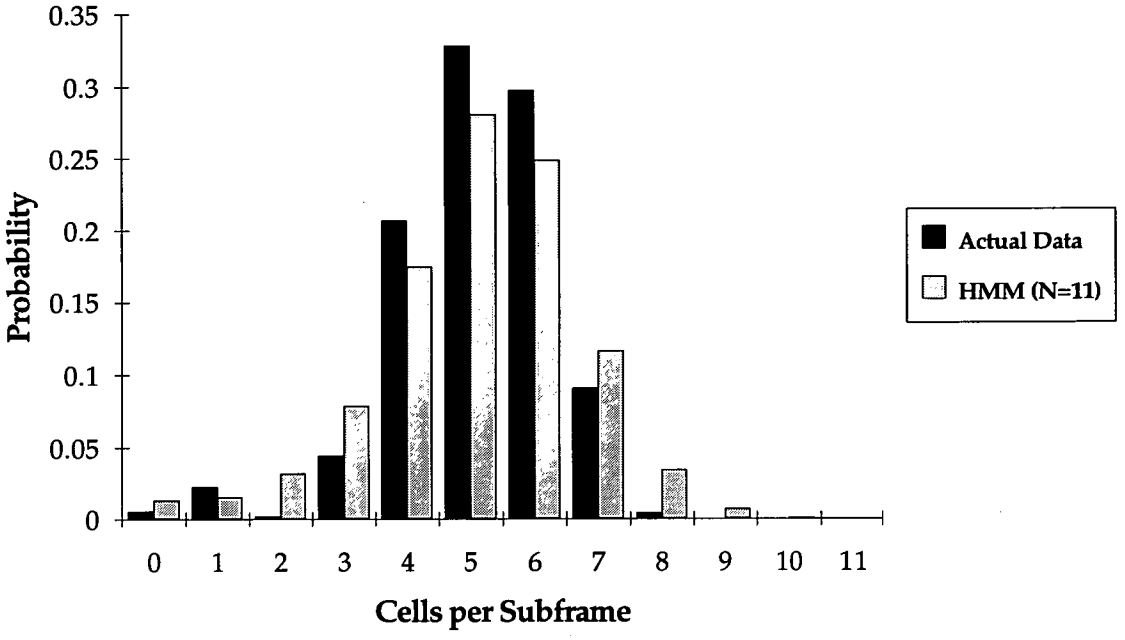


Figure 6.6: Pdf of the number of cells per subframe for N=11 of the actual data and HMM data

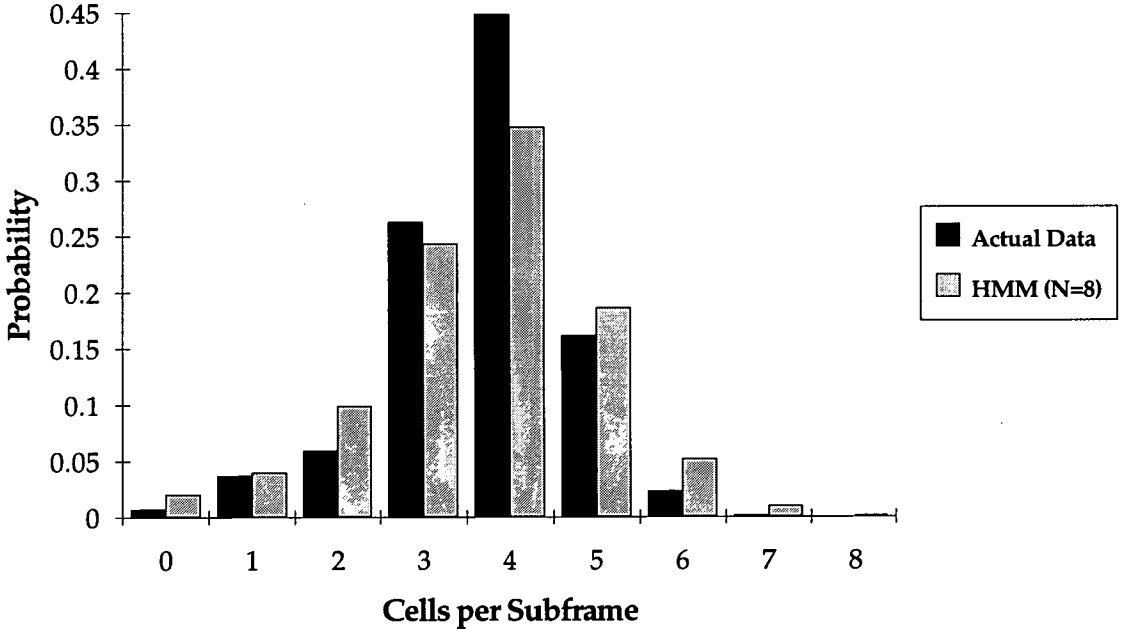


Figure 6.7: Pdf of the number of cells per subframe for N=8 of the actual data and HMM data

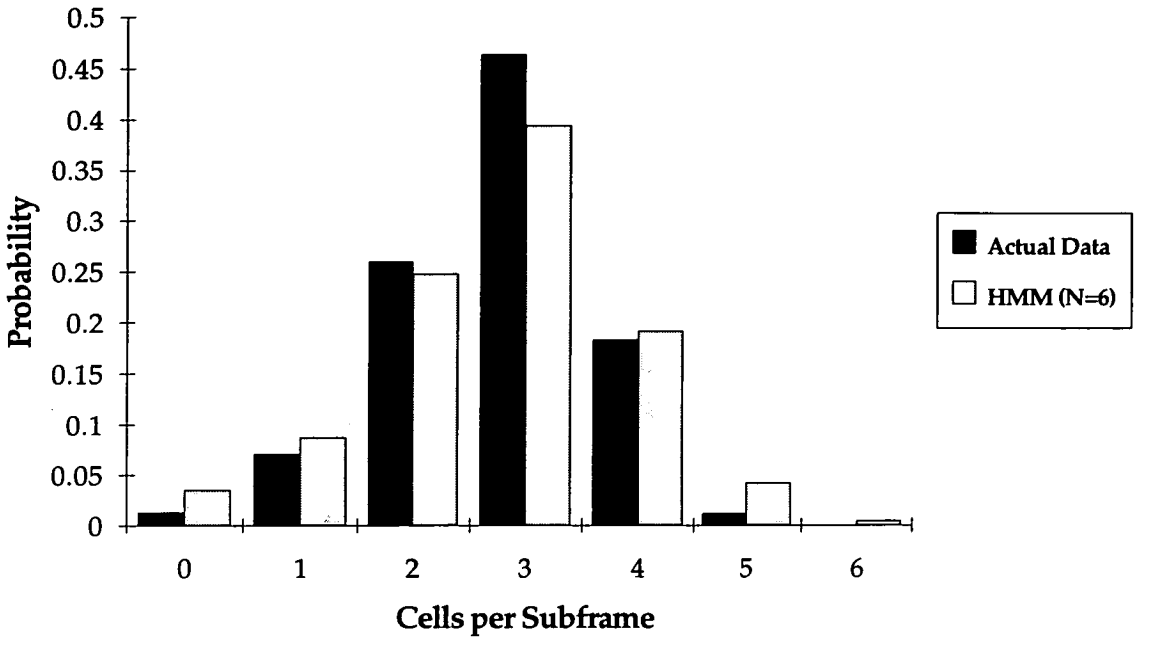


Figure 6.8: Pdf of the number of cells per subframe for N=6 of the actual data and HMM data

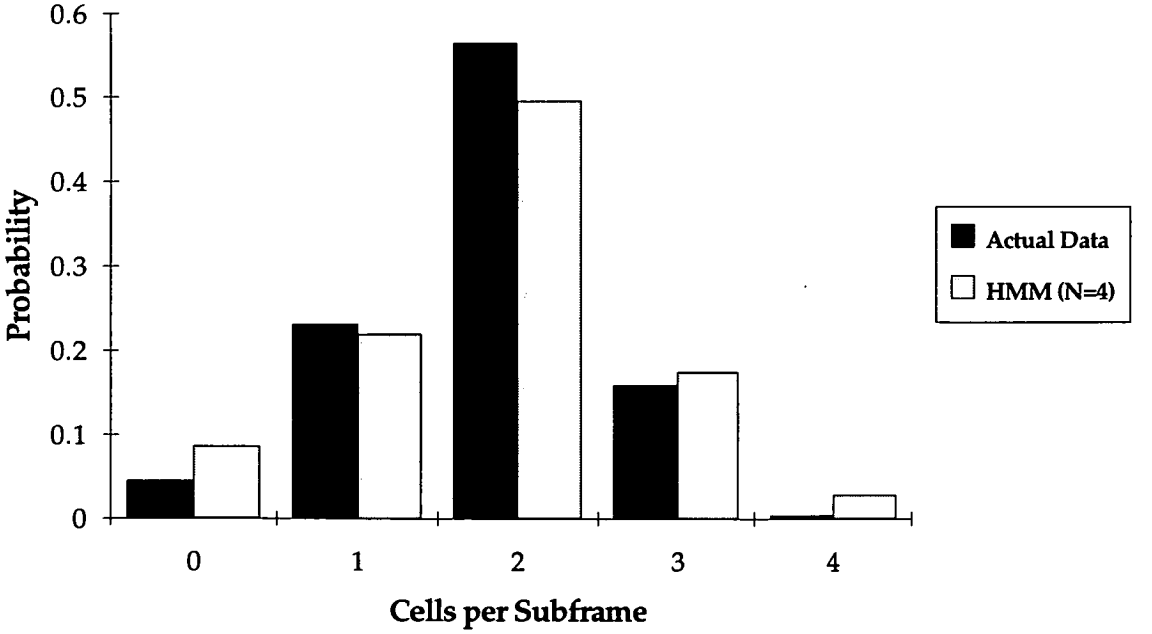


Figure 6.9: Pdf of the number of cells per subframe for N=4 of the actual data and HMM data

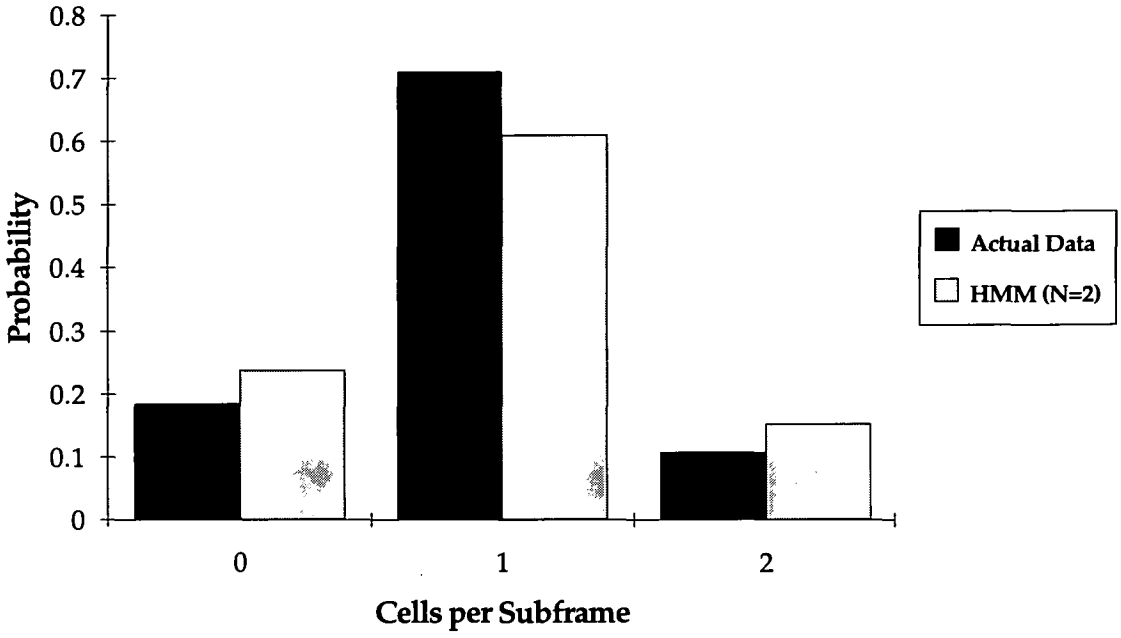


Figure 6.10: Pdf of the number of cells per subframe for N=2 of the actual data and HMM data

However, while these probability density functions suggest the HMM to be a promising model for the VBR video under consideration, it is not obvious how good these models will be for queueing purposes and what subframe size would give the closest HMM fit to the actual video data. It may be logical to think that the larger the number of the modes, the better the fit it should give. Figures 6.11 and 6.12 however do not support this argument. These figures show the mean queue length and the standard deviation of queue length when each of these traffic is fed into a dedicated server. These results have been obtained by simulating each traffic model for 1 hour.

It appears from Figures 6.11 and 6.12 that for utilisations up to about 70%, all models give an excellent fit to the actual video data. However for the highest utilisation undertaken (i.e. 90%), it is obvious that larger subframe sizes (and larger number of modes) have resulted in worse fits as compared to the smaller subframe sizes. It can be seen that for N=4, even for 90% utilisation a very good fit is obtained both for server queue length and for the standard deviation of the queue length.

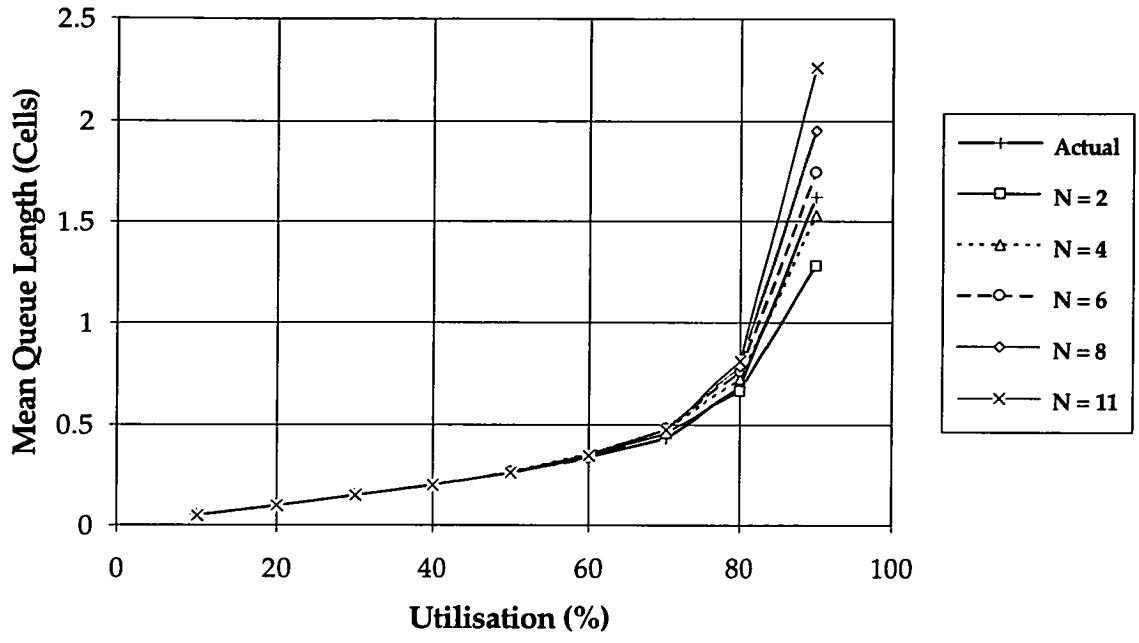


Figure 6.11: HMM mean queue size for various values of N

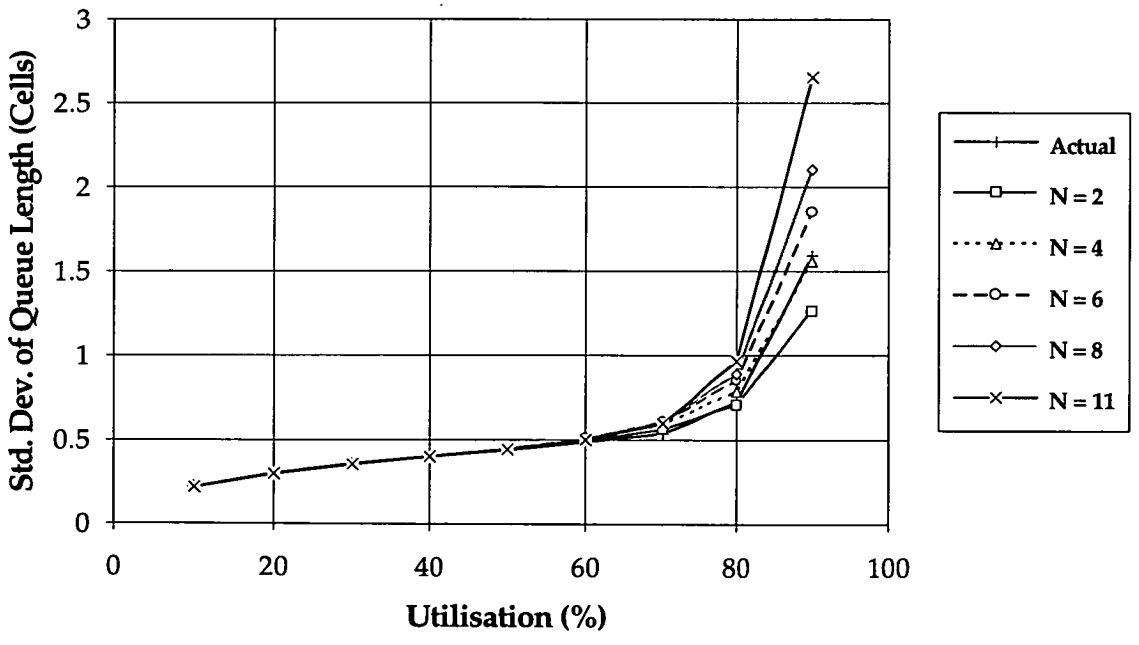


Figure 6.12: HMM standard deviation of queue size for various values of N

Figures 6.13 and 6.14 show convergences of mean queue length and standard deviation of queue length as a function of time for $N = 11$ and 90% utilisation (worst fit observed). These indicate that the difference between the worst fit and the actual video data is not related to the length of run of the simulation. In fact, it can be seen that these results converge quickly.

It is appropriate at this stage to undertake a correlation study of the ATM cell streams resulting from the actual video and from the best fit and the worst fit models (i.e. $N = 4$ and $N = 11$). Mathematical details of these correlation analyses are shown in Appendix B. Figure 6.15 shows $\hat{R}_{xy}(m)$, the normalised auto-correlation of cell/block generation of the actual video data for a block *offset* of up to 1200. The results have been obtained by processing 449 frames of the Salesman sequence. It can be seen that there is a strong correlation present at the frame rate (1 frame = 396 blocks). This is to be expected because there are a lot of similarities between blocks of consecutive frames. Let us present the numerical values of the three peaks shown in Figure 6.15 to show how fast this correlation dies out. With an offset equal to 396 blocks (size of 1 frame in blocks) the auto-correlation is 0.2510, and for offsets of 792 and 1188 blocks (sizes of 2 and 3 frames respectively) the auto-correlation drops to 0.2479 and 0.2431 respectively. Figure 6.16 shows how the correlation related to frame rate dies out as a function of the number of frames.

Figure 6.17 is a close-up of Figure 6.15. From that figure we can interpret that for the sequence under study, if one block generates a cell, the next block is unlikely to generate another cell, but the block after that will most likely generate one. The noticeable correlation around offset=22 is related to a line period. Note that for this sequence, there are 352 pixels per line and we have a total of 288 lines. Therefore with blocks of 16×16 pixels, each frame consists of 18 rows of 22 blocks each.

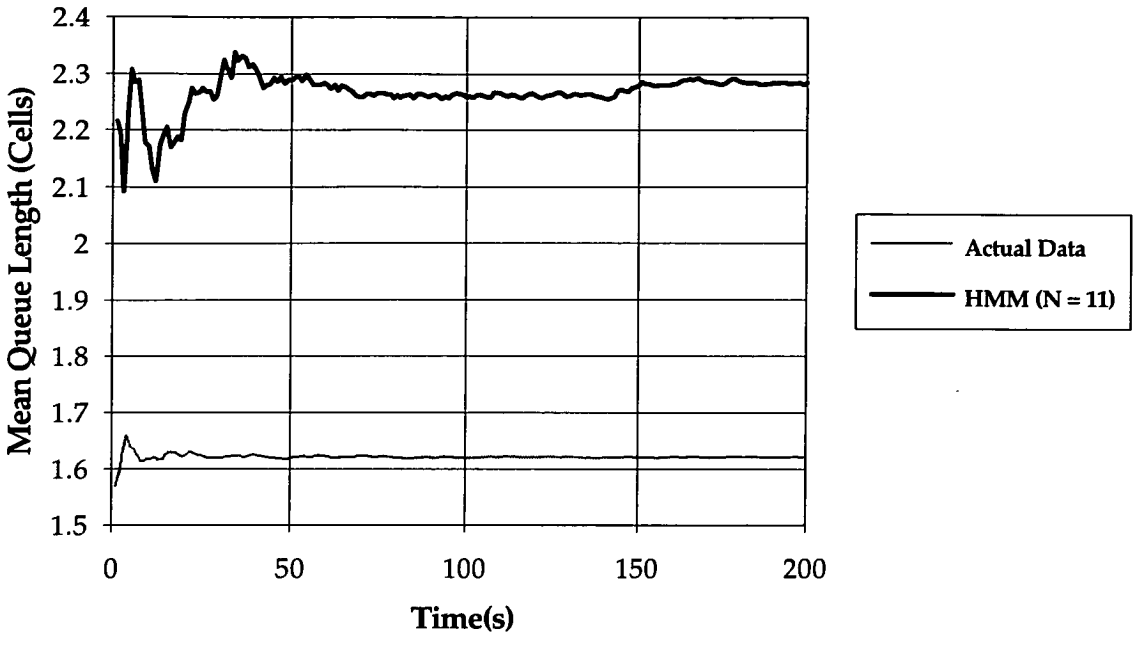


Figure 6.13: HMM mean queue length as a function of time

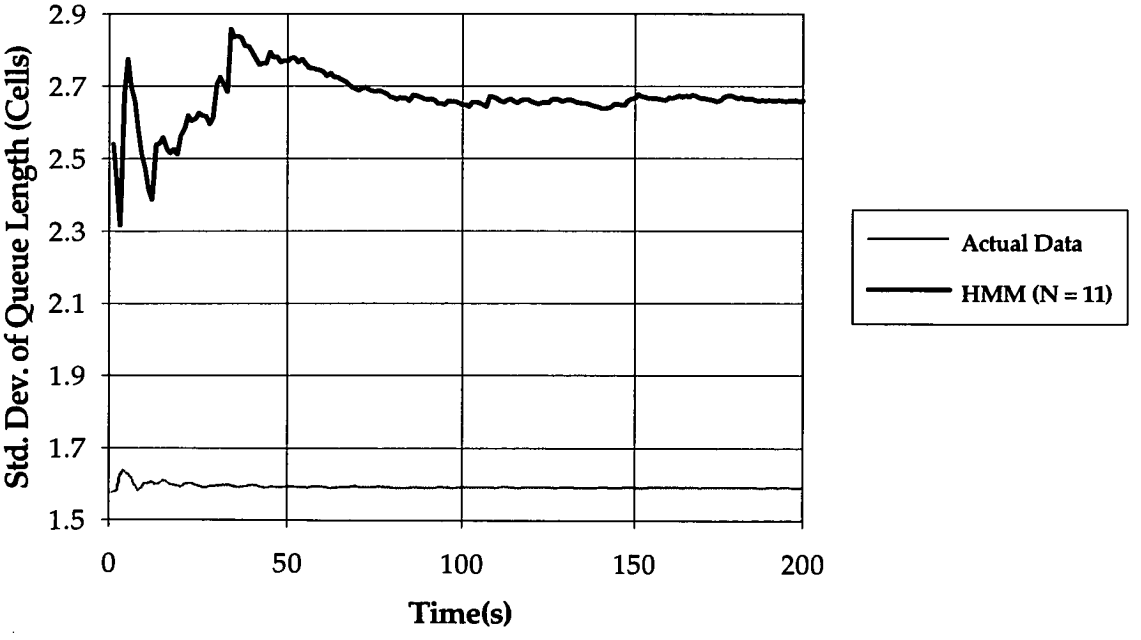


Figure 6.14: HMM standard deviation of queue length as a function of time

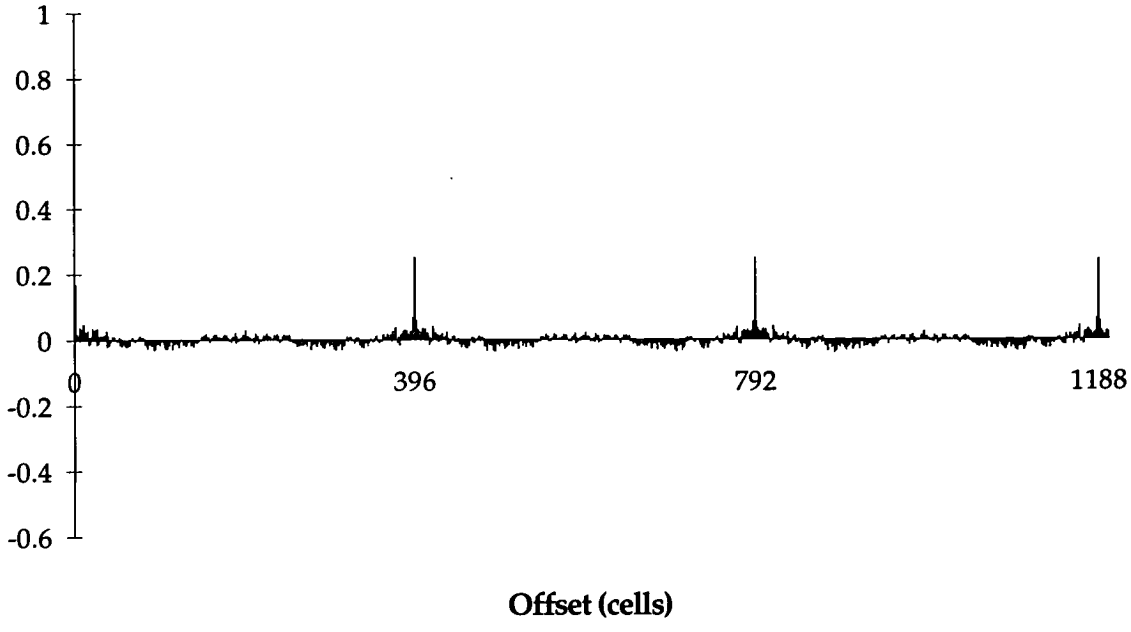


Figure 6.15: Normalised autocorrelation of cell/block generation of actual data for the Salesman sequence

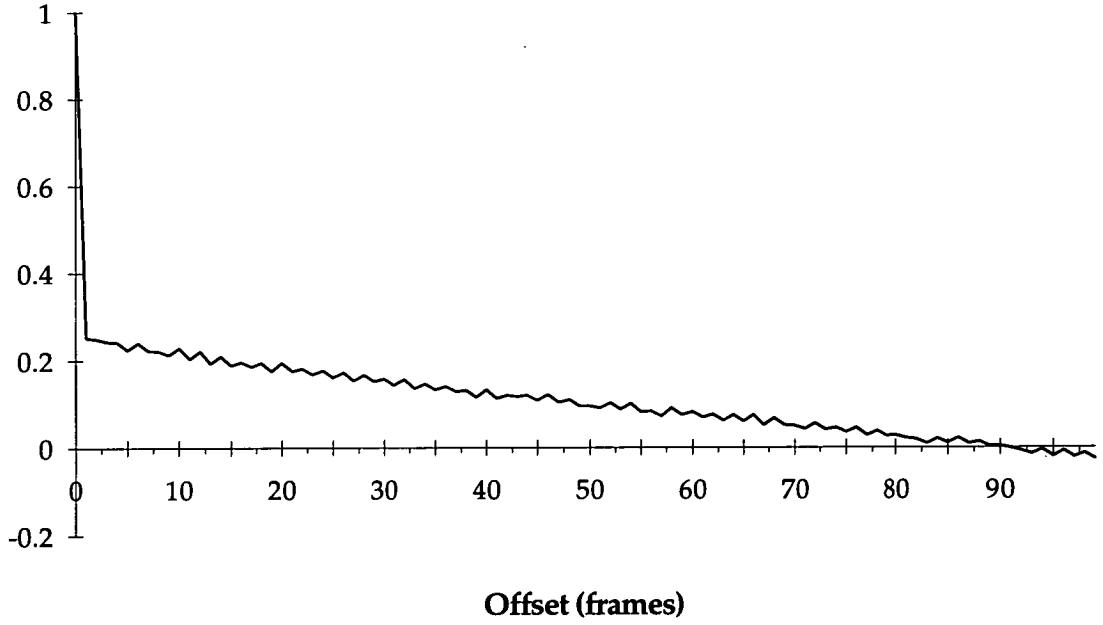


Figure 6.16: Normalised autocorrelation of cell/block generation of actual data for the Salesman sequence on a frame to frame basis

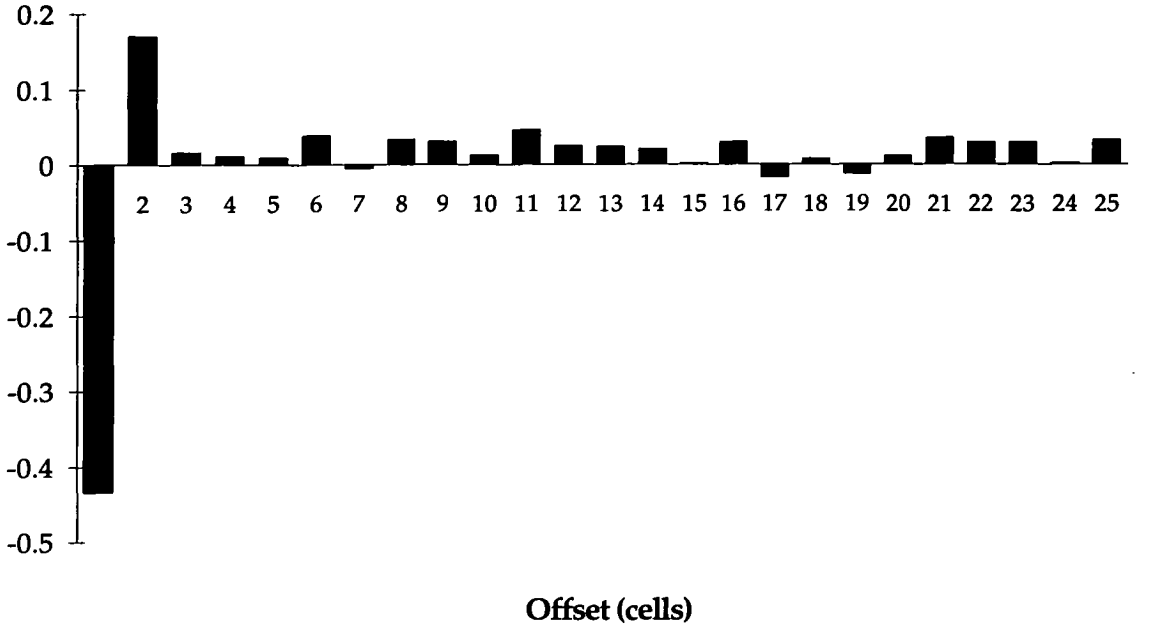


Figure 6.17: Normalised autocorrelation of cell/block generation of actual data for the Salesman sequence

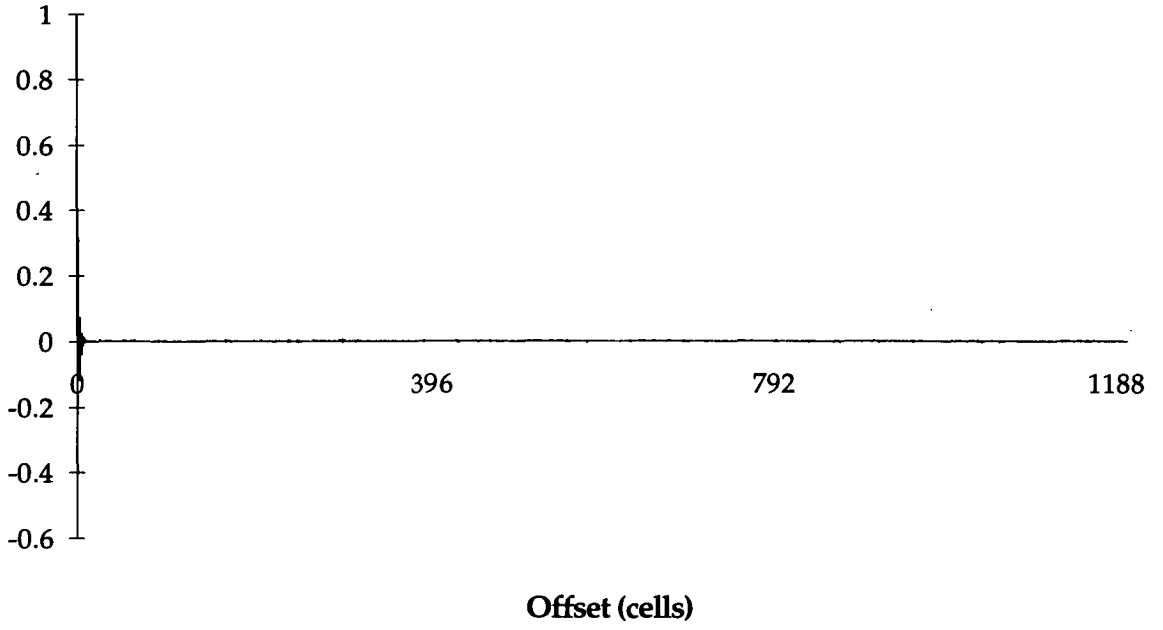


Figure 6.18: Normalised autocorrelation of cell/block generation of HMM data (N=4) for the Salesman sequence

Figure 6.18 shows the autocorrelation for HMM data with N=4. It is obvious that the correlation at the frame rate does not exist and, in general, there is only

very short term correlation present. Now let us compare the short term correlation present in the best fit (N=4) and worst fit (N=11) to that of the actual data. These are shown in Figures 6.19 and 6.20. It was hoped that these figures would give some more insight as to why one case should give a better fit to the actual data than the other, but, comparison of the two figures does not provide an answer.

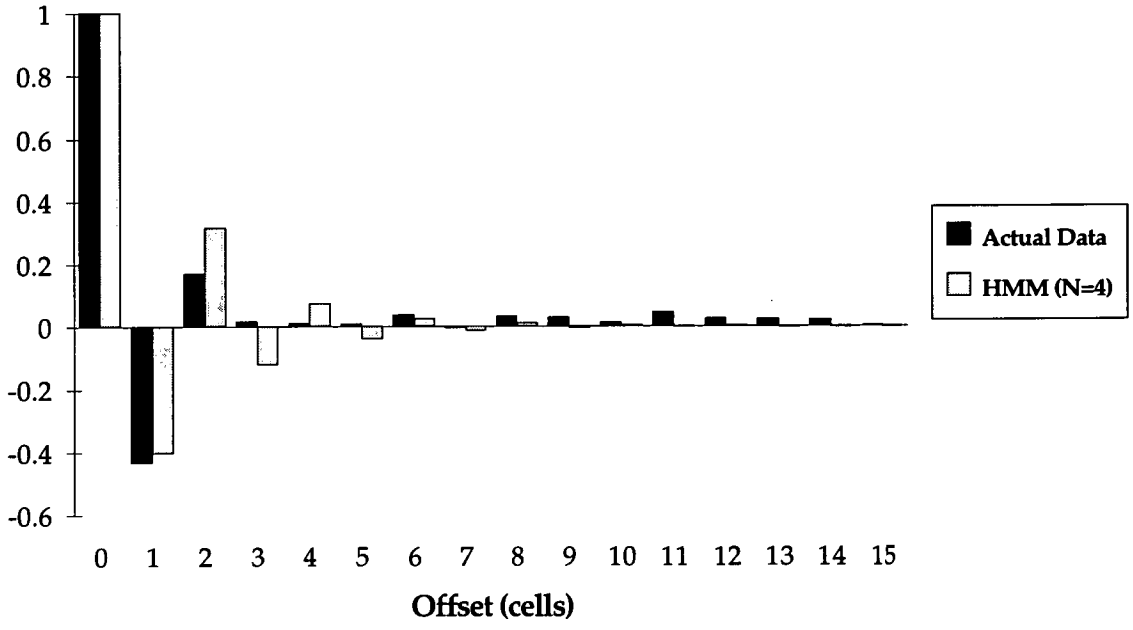


Figure 6.19: Normalised autocorrelation of cell/block generation of actual data and HMM data (N=4) for the Salesman sequence

At this stage, we shall reconsider the Hidden Markov Model (HMM) and look for ways of improving and/or simplifying it. From the pdf plots (Figures 6.6 to 6.10), it can be seen that the probability density functions of the number of cells generated in a subframe are not peaky enough (compared to the actual video) in the middle regions, and that they are higher than those of actual video data at the tails of the functions.

This phenomena is due to the Markov models used for cell generation within individual modes. Note that when assigning modes to the traffic generated from the actual video, we identify a mode by the number of cells generated from a

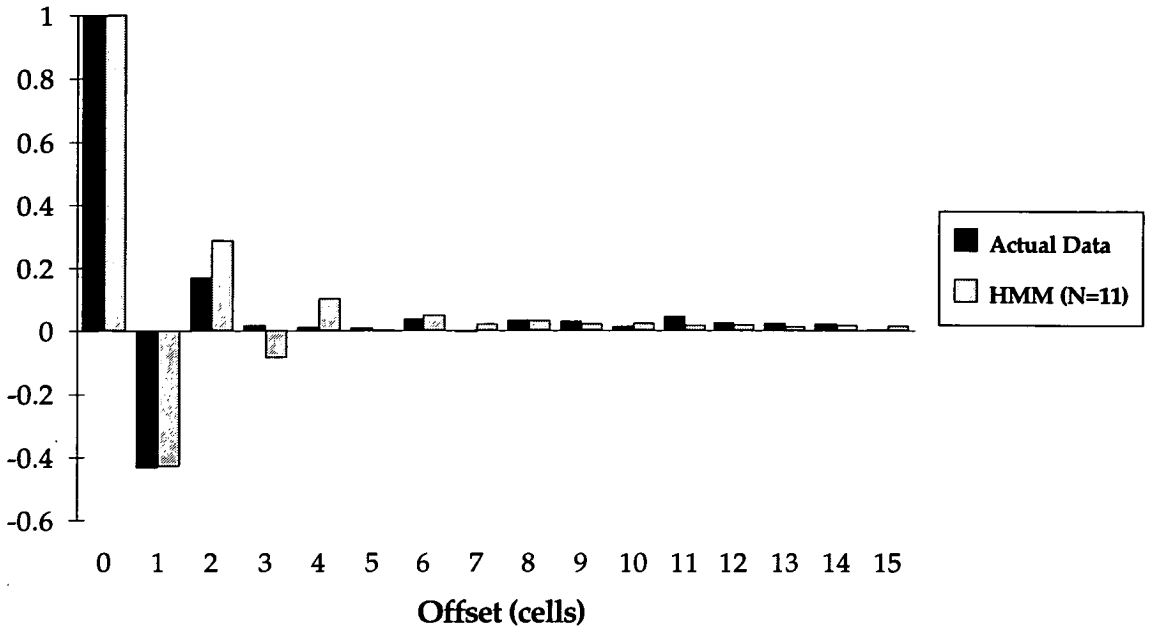


Figure 6.20: Normalised autocorrelation of cell/block generation of actual data and HMM data (N=11) for the Salesman sequence

subframe (see Table 6.3). These mode identifications are used for generating the intermode transition probabilities matrix, P. From the pattern of cell generation within subframes belonging to a particular mode, we generate the intramode cell generation transition probabilities matrices, $A_0 \cdots A_{m-1}$. These matrices are then used to generate the HMM traffic.

The important point to consider here is that from the number of cells generated in a subframe of HMM traffic, we cannot precisely identify the associated mode. As an example consider the case of $N = 8$ in Table 6.3. For the actual data there, every time the number of cells generated from a subframe is 4, we can confidently say that the video traffic is in mode 4. However, when looking at the traffic generated from the HMM, if the number of cells generated from a subframe is 4, we can say that *most likely* we are in mode 4. This is because for the HMM data it is also possible that 4 cells be generated in modes 3 or 5 and with a lesser likelihood in higher and lower modes. This is due to random nature of the Markov models used to specify cell generation patterns in particular modes. The question to be asked is: what is more important to the accuracy of the overall

model; the exact number of cells in each mode, or the correlation between cells generated in a particular mode? In the next section we shall examine a model that guarantees the generation of the exact number of cells for each mode, but disregards the cell generation transition probabilities in the consecutive blocks of a subframe.

6.4 HMD: HMM with Deterministic number of cells in each mode

This model is different from the HMM in that Markov models are no longer used for cell generation within a mode. This model is fully specified only by the inter-mode transition probabilities matrix and knowledge of mode assignments. In the last section we explained why the pdfs of HMM were less peaky in the middle regions and peakier at the tails in comparison with the actual video traffic. The purpose of designing the HMD model is to force the pdf of cells per subframe of the model traffic to closely follow that of the actual video traffic.

We need to ensure that if k cells are generated per subframe in mode i ($i = 1 \cdots m$) for real data , then when in mode i of the model, exactly k cells are generated during that subframe. This is achieved by normalising the size of the subframe to 1 (as shown in Figure 6.21) and generating k random numbers in the range [0,1), the values of which will determine the blocks from which each of the k cells is generated.

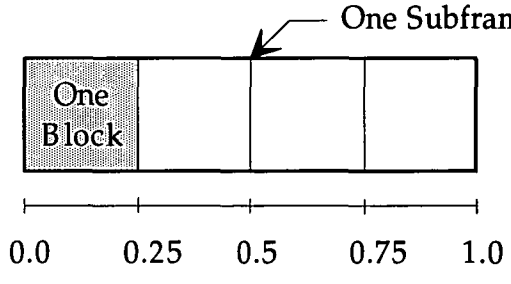


Figure 6.21: Normalising the size of the subframe to 1 ($N = 4$)

	N=8		*N=6*		*N=4*		*N=2*	
Cells per subframe	*Mode index*	*Prob.*	*Mode index*	*Prob.*	*Mode Index*	*Prob.*	*Mode Index*	*Prob.*
0	0	0.00715	0	0.01262	0	0.04472	0	0.18383
1	1	0.03689	1	0.07046	1	0.23081	1	0.70974
2	2	0.05921	2	0.25912	2	0.56284	2	0.10642
3	3	0.26276	3	0.46392	3	0.15779		0
4	4	0.44835	4	0.18218	4	0.00382		
5	5	0.16161	5	0.01154		0		
6	6	0.02272	6	0.00013				
7	7	0.00126		0				
8		0						

Table 6.4: HMD mode assignment for various values of N

One of the implications of this model is that the mapping of the number of cells generated from a subframe to a particular mode must preferably be a one to one mapping, i.e. it is better not to combine states which correspond to different number of cells per subframe into one mode. Table 6.4 shows the mode assignment for the HMD model.

This model was investigated for subframe sizes of 2, 4, 6 and 8. For the HMD data, the resulting pdfs of the number of cells in a subframe is almost identical to the actual video data. Two of these are shown in Figures 6.22 and 6.23. However, comparing Figures 6.24 and 6.25, the server mean queue length and the standard deviation of the queue length for the actual video data and the HMD data reveals that the quality of the fit has deteriorated. The queueing results of the models now diverge from the results of the actual video data at much lower utilisations. The autocorrelation analysis of this model for N=4 (see Figure 6.26) shows that the short term correlation of the HMD data is very different from that of the actual video data.

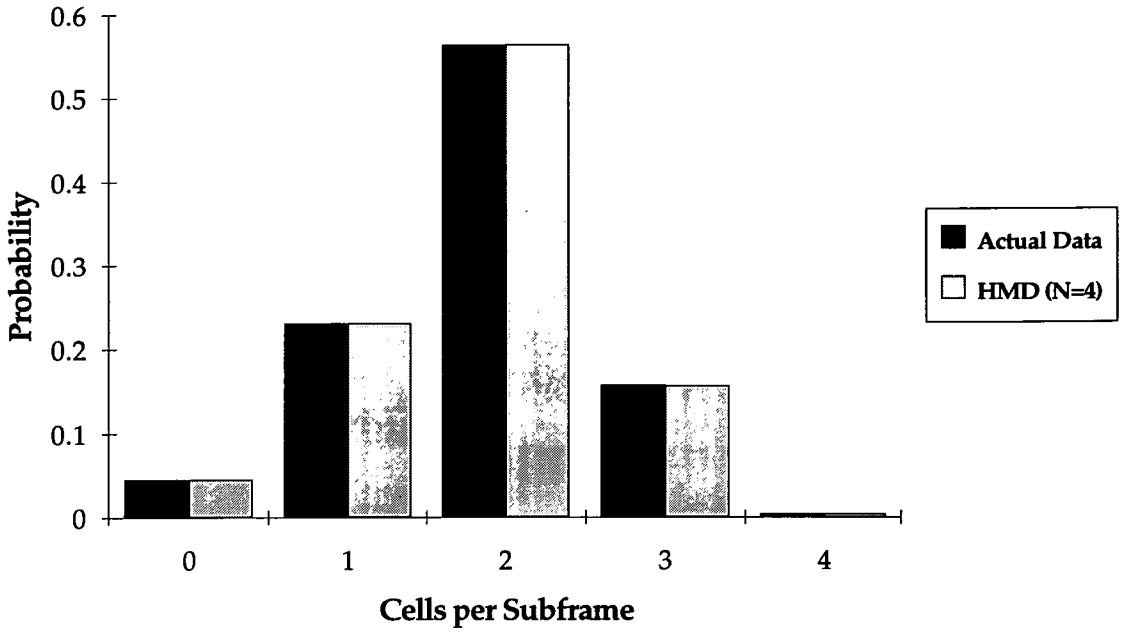


Figure 6.22: Pdf of the number of cells per subframe for N=4 of the actual data and the HMD data

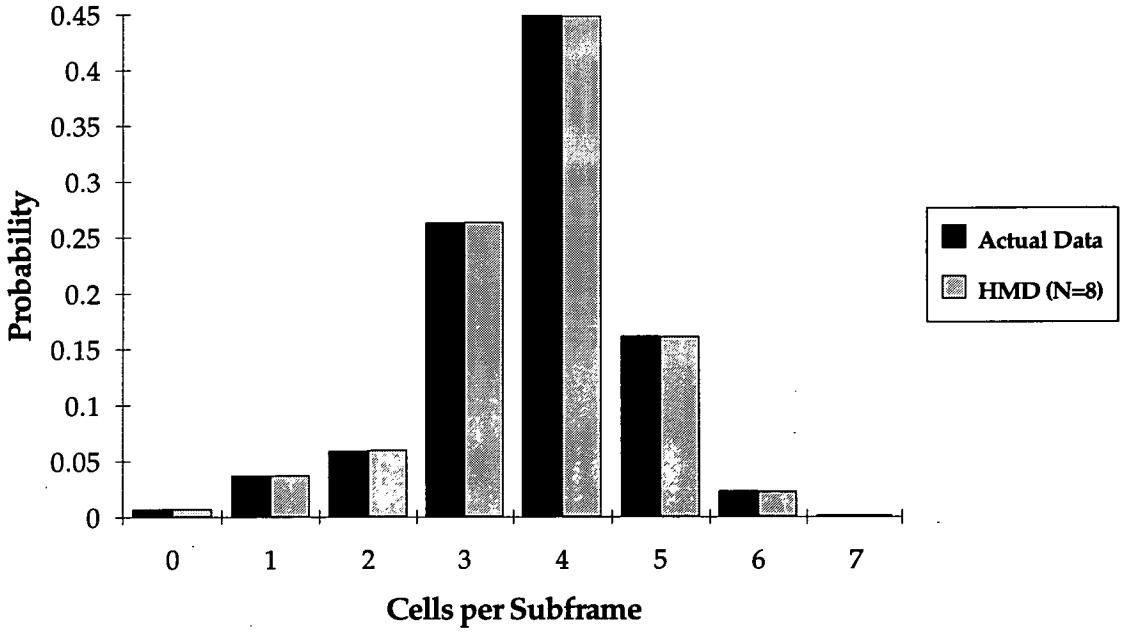


Figure 6.23: Pdf of the number of cells per subframe for N=8 of the actual data and the HMD data

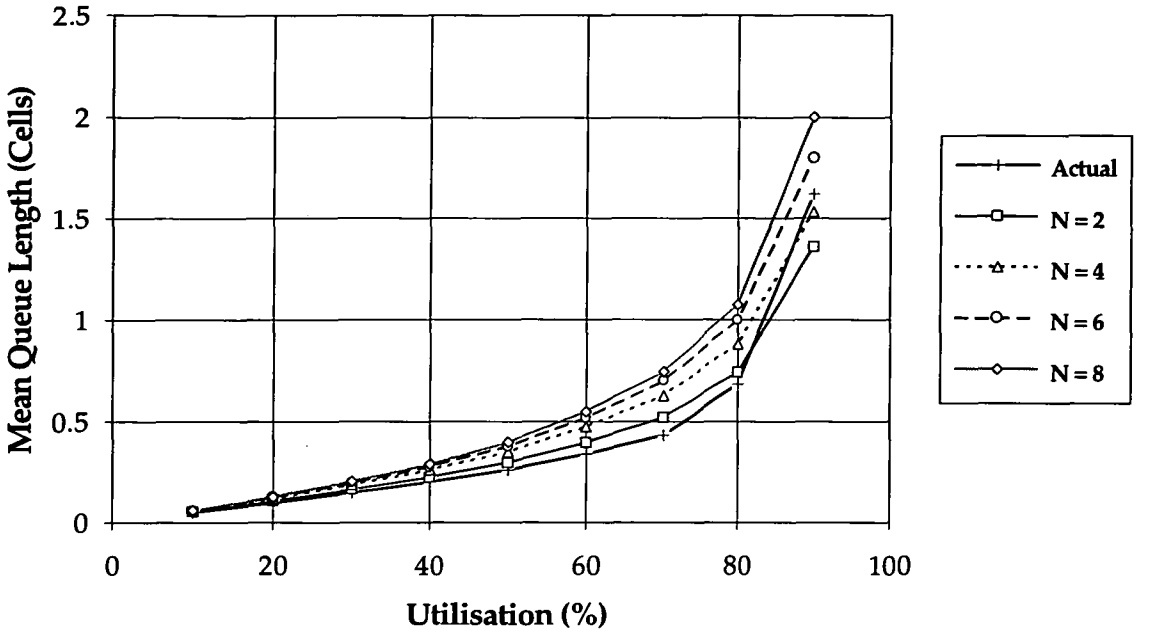


Figure 6.24: HMD mean queue size for various values of N

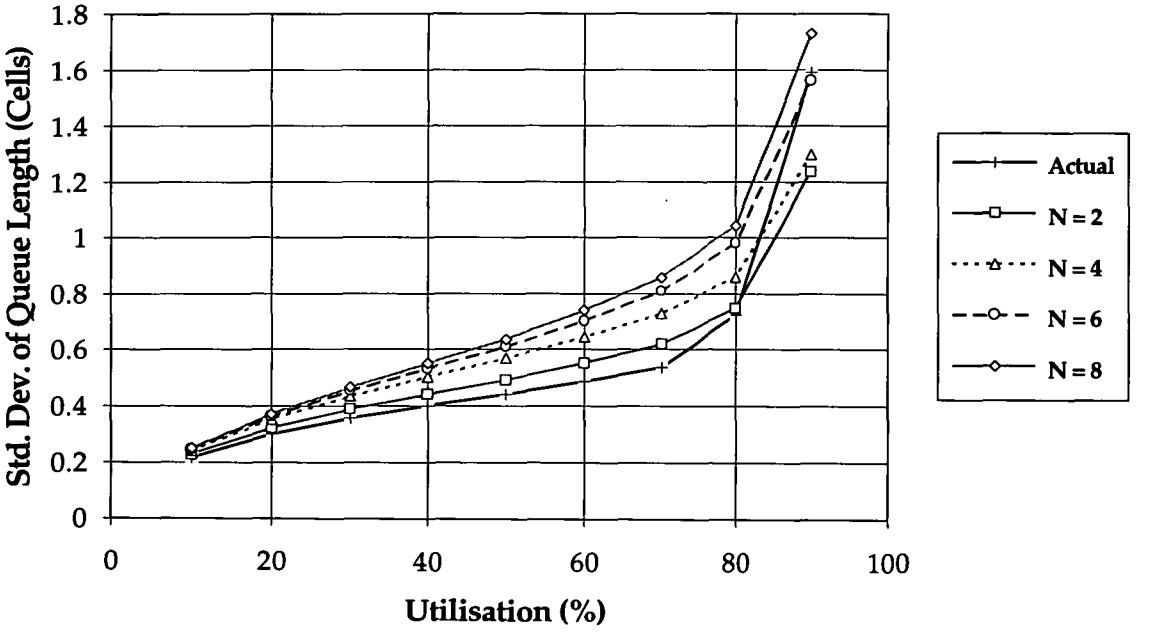


Figure 6.25: HMD standard deviation of queue size for various values of N

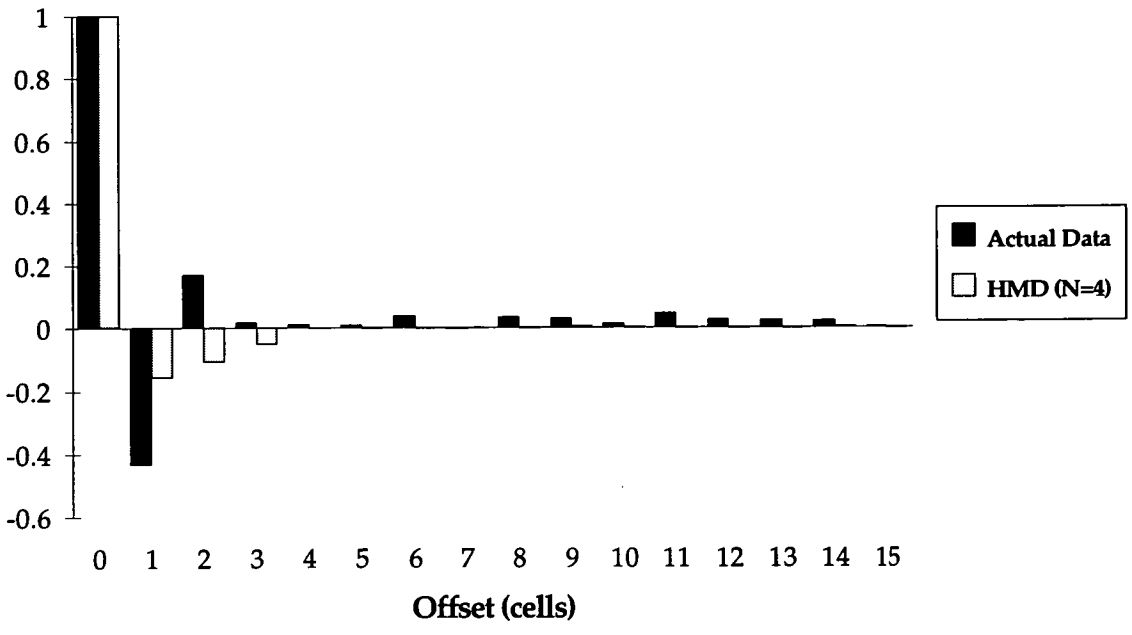


Figure 6.26: Normalised autocorrelation of cell/block generation of the actual data and the HMD data (N=4) for the Salesman sequence

If we ponder upon the definition of the HMD model, we realise one of the main sources of error. The error could arise from the fact that in the HMD model, while generating k cells in a subframe in mode m, it is possible to get some blocks in that subframe that generate more cells than the maximum cells per block observed in the actual video data. In fact, in the HMD model it is even possible for all the k cells in the subframe to end up on the same block. For example, the Salesman sequence after compression can only give a maximum of 1 ATM cell per block. This maximum value is much larger for the HMD data, depending on the maximum number of cells that can be generated in the highest bit rate mode. In the next section we describe how the HMD model can be improved for a better fit to the actual video data.

6.5 HMDL: HMD with Limited cells/block

The HMD model can be modified to include limits on the maximum number of cells per block. This model is called HMDL. Each mode will still generate a deterministic number of cells in a subframe, but we do not allow any block to carry

more cells than the maximum cells per block of the actual video data.

If for real data k cells are generated per subframe in mode m, then when in mode m of the model, exactly k cells are generated during that subframe. As with the HMD case, the size of the subframe is normalised to 1. A random number is generated in the range [0,1) the value of which will correspond to a particular block of the subframe. However, before assigning a cell to that block, the number of cells already assigned to that block is checked. If this number has not reached the maximum cells per block of the actual video data, the assignment is carried, otherwise, the random number is discarded and another one is generated. This process is repeated until the target number of cells in the current subframe is reached.

The mode assignment table for the HMDL is identical to the HMD case (see Table 6.4). The resulting pdfs of cells per subframe are also the same as the HMD case. Figures 6.27 and 6.28 show that in the queueing results there is a significant improvement in the fit of the model to the actual video data compared to the HMD model. Any discrepancy left is due to disregarding the pattern of cell generation in various modes of the actual video data. This, however, does not seem to have a significant effect on the accuracy of the models for queueing purposes.

6.6 Summary

In this chapter we have shown that hidden Markov models can successfully be applied in the modelling of variable bit rate video services. From the original HMM we arrived at a simplified version, the HMDL model, which requires much fewer number of parameters for its complete specification. Although the HMDL is much simpler than the HMM, it can still track the original video data very closely for queueing purposes.

These video traffic models were based on dividing each video frame into several

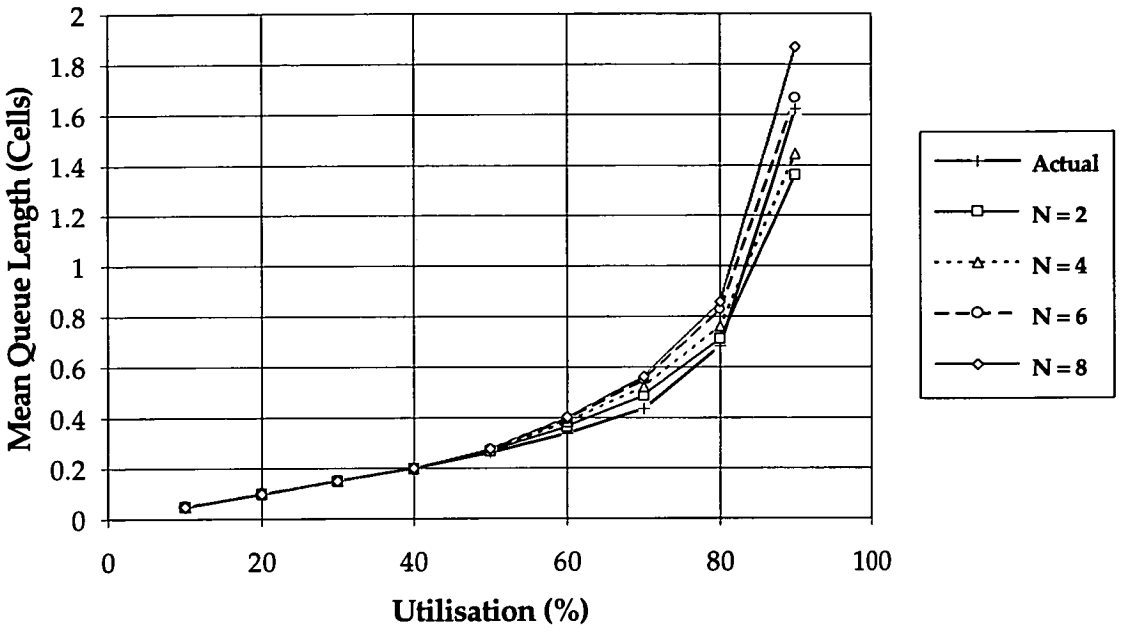


Figure 6.27: HMDL mean queue size for various values of N

fixed size blocks, then grouping a number of blocks into a subframe and assigning a mode to each subframe depending on the number of ATM cells generated from that subframe. The original HMM for video traffic consisted of an intermode transition probability matrix for modelling the transitions between various modes, and several (as many as the number of modes) intramode cell generation transition probability matrices for modelling the cell generation within individual modes. Each traffic (from a model or from the actual video data) was fed into a single server queue. Some results were generated for the mean and the variance of the queue population. Probability density functions (pdfs) for the number of ATM cells per subframe were also generated. The results of the models for various sizes of subframe were compared with those from the actual video traffic. This comparison indicated that the size of the subframe had a significant effect on the accuracy of the model, particularly at high utilisations.

We also investigated the HMD model which is a simplified version of the original HMM model, where intramode matrices are no longer used to model the cell generation within each mode. Instead a deterministic number of cells are generated in each mode, randomly distributed over the blocks of the subframe. The pdf

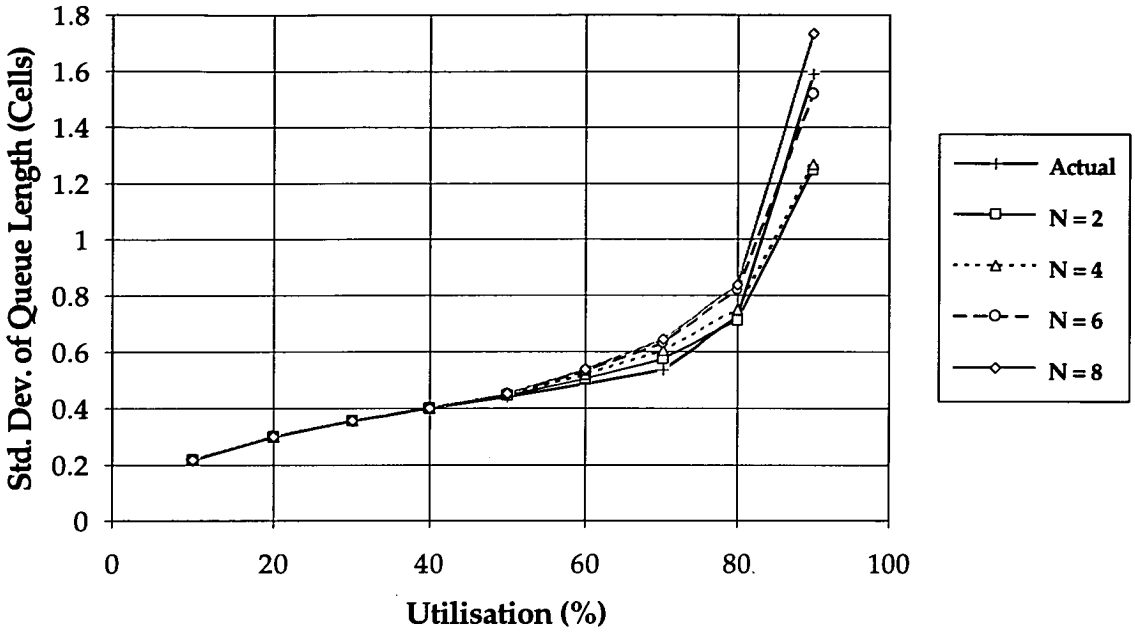


Figure 6.28: HMDL standard deviation of queue size for various values of N

results for the HMD model showed great improvement over the pdf results for the original HMM model. In fact, for the HMD, the pdf results were almost identical to those of the actual video traffic. However, for high utilisations some loss of accuracy had been incurred in the results for the mean and the variance of the queue population.

The next model investigated was HMDL, which was similar to HMD except that no block within a subframe could generate more ATM cells than the maximum observed value for the real video traffic. This resulted in a significant gain in the accuracy of the model.

Correlation studies of the video traffic indicated that strong cyclic variations were present in the cell stream of the video traffic, particularly at the frame rate and at the line rate.

As for the size of the subframe, it is clear that it is a parameter subject to optimisation. The best value of the subframe size would depend on the rate of variation in local complexity of the blocks in one row. To fully understand the

effect of subframe size on the accuracy of the model, it is necessary to repeat the processes described in these models for a wide range of video services subject to different levels of compression. However, we have had available to us only one video coding program and therefore we have not been able to try these models for video traffic subjected to other levels of compression. Research in the area of video coding and compression is not within the scope this thesis.

Chapter 7

Queues with Periodic Arrival Rates

7.1 Introduction

In the studies of variable bit rate video presented in Chapter 6 and published by Habibi in [118], it has been shown that the pattern of cell generation from a variable bit rate video codec is strongly correlated (e.g. at the frame rate, the line rate, etc.). The periodic variations in the arrival rate of such services prompted an investigation of analytical methods that can cater for such periodicities. A summary of the research outlined in this chapter and in the next chapter has been published by Habibi et al in [129] and [130].

In this chapter, the emphasis is shifted from a classical queueing problem where the arrivals are random with a particular mean value, to a system where arrivals are still random, but the mean arrival rate varies periodically. It is shown how Fourier series can be used as a tool to analyse such a system. One example of cyclic arrivals is the traffic generated by a video signal that has been subjected to incomplete redundancy removal. Also when several similar sources are multiplexed in a network, because all these sources use the network clock for synchronisation, the multiplexed traffic will show cyclic patterns. Another example where cyclic traffic may be generated is at a multiplexer with its input streams having slightly different rate variations (beating effect). This technique could also

be applied to the queueing analysis of PSTNs where the arrival of calls tend to have a periodic pattern, i.e. the number of call requests depends on the time of the day, the day of the week and perhaps the time of the year. For the case of a PSTN, the period can be taken as one week, or, it could be extended to one year if enough statistics are available and if there is enough computer memory to handle the size of the problem. The advantage of this technique is that by solving the problem for the particular queueing system only once, all the usual performance parameters can be calculated as a function of time for the whole cycle. For example, in a PSTN the blocking probability of calls can be obtained as a function of time for the whole period.

While the exact problem and the analysis techniques presented in this chapter and in the next chapter are unique, some related work may be found in the literature. For example Latouche [131] considers a packet system where packets are submitted from voice sources which are intermittently active and all have the same input rates. He asserts that over short intervals of time, circuit emulation type sources submit packets in a deterministic and periodic manner. Latouche [131] refers to the results of [132] and concludes that periodic cycles in the packet arrival process induce periodic cycles in the buffer process. The approach of [131] to this problem is fairly theoretical and it does not present any quantitative results for the contents of the buffer. In this chapter however, we will outline a method for quantifying the content of the buffer in similar situations.

Most studies in the area of cyclic queues have been focused on cyclic service systems [133, 134, 135, 136, 137, 138]. These systems are also known as polling systems or token-passing systems. In [139], Leung analyses an asymmetric cyclic-service system with a probabilistically-limited (P-L) service policy. The P-L service policy means that during each visit of the server to a queue, the maximum number of customers served is determined by a probability which is independent of system states and can be different for various queues. A visit is defined as the period of time in which a queue is being continuously served by the server. Leung uses a numerical technique based on discrete Fourier transforms (DFTs) to solve for the distributions of queue population. The queue length distribu-

tions are expressed as functions of the probability generating functions (pgf's) for the state probabilities observed at visit-completion instants. The response time and waiting time distributions are obtained from the probability generating functions. These probability generating functions are solved using an iterative numerical technique based on DFTs.

Another relevant work is by Levy et al [140] where they consider the problem of calculating the expected delay in polling systems with zero length switch-over periods. The models considered in [140] consist of N infinite capacity independent queues that are served by a single server. The system switches over between these queues by a set of rules that define the polling order. The polling orders considered by [140] are: *cyclic* - where after serving queue i, the server switches over to queue $i+1$ (modulo N), and *memoryless random* - where the next queue to be served is queue j with probability p_j.

7.2 A Simple Queueing System

Let us start with the simple case of a single server queue with random arrival and random service rates and from there proceed to the case of periodic arrival rates. Let n denote the state of the system where n is the population of the system at time t. Let $\lambda_n(t)$ be the arrival rate and $\mu_n(t)$ be the service rate of the system in state n and at time t. Let $P_n(t)$ be the probability of having n traffic entities in the system at time t. Let Δt be an infinitesimally small time interval so that the probability of more than one event occurring during time interval Δt is zero. Obviously a negative population is not possible. Furthermore, there is no service when the population of the system is zero. Therefore for $n \geq 0$, the probability of being in state n at time $t + \Delta t$ can be written as

$$\begin{aligned} P_n(t+\Delta t) &= \{1 - \lambda_n(t)\Delta t - \mu_n(t)\Delta t\}P_n(t) + \lambda_{n-1}(t)\Delta t P_{n-1}(t) \\ &\quad +\mu_{n+1}(t)\Delta t P_{n+1}(t) \ . \end{aligned} \tag{7.1}$$

For $n = 0$ this equation reduces to

$$P_n(t+\Delta t) = \{1 - \lambda_n(t)\Delta t\}P_n(t) + \mu_{n+1}(t)\Delta t P_{n+1}(t) \ . \tag{7.2}$$

For $n > 0$, differentiating equation (7.1) with respect to time gives

$$\frac{dP_n(t)}{dt} = -\{\lambda_n(t) + \mu_n(t)\}P_n(t) + \lambda_{n-1}(t)P_{n-1}(t) + \mu_{n+1}(t)P_{n+1}(t) \tag{7.3}$$

and for $n = 0$ the corresponding equation is

$$\frac{dP_n(t)}{dt} = -\{\lambda_n(t)\}P_n(t) + \mu_{n+1}(t)P_{n+1}(t) \ . \tag{7.4}$$

7.3 The M/M/1 Queueing System

An M/M/1 system is a special case of the system described in the last section with the arrival rate and the service rate independent of n (nevertheless, for $n = 0$, the service rate is zero). For this case the n subscripts are dropped from equations (7.1) & (7.2), therefore for $n > 0$ we have

$$\frac{dP_n(t)}{dt} = -\{\lambda(t) + \mu(t)\}P_n(t) + \lambda(t)P_{n-1}(t) + \mu(t)P_{n+1}(t) \tag{7.5}$$

and for $n = 0$ we have

$$\frac{dP_n(t)}{dt} = -\lambda(t)P_n(t) + \mu(t)P_{n+1}(t) \ . \tag{7.6}$$

7.4 Sinusoidally Varying Mean Arrival Rate

While sinusoidal patterns of arrival rates are not usual, they simplify the development of a solution technique for cyclic arrivals. After a solution to this simplified case is found, the analysis will be generalised to cater for arbitrary shapes of arrival rate. Here, the arrivals are assumed to be random, with a mean value that oscillates sinusoidally around a fixed term, i.e.

$$\lambda(t) = \alpha + \beta \sin \omega t \ . \tag{7.7}$$

Equation 7.7 implies that $\beta \leq \alpha$ because negative arrivals are not possible. The service completion time is also taken to be random, but with a mean service rate that is not a function of time, i.e. $\mu(t) = \mu$. The subscript n is reintroduced for the service rate so that the special case of $n = 0$ does not require separate treatment:

$$\mu_n = \begin{cases} \mu, & n > 0 \\ 0, & n = 0 \end{cases} . \tag{7.8}$$

With these provisions equation (7.3) becomes:

$$\frac{dP_n(t)}{dt} = -\{\lambda_n(t) + \mu_n\}P_n(t) + \lambda_{n-1}(t)P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) \ . \tag{7.9}$$

We conjecture that if $\lambda(t)$ is periodic and the queue is stable, then the probabilities of being in various states will, in the long run, acquire the same periodicity as the arrival rate. This implies that $P_n(t)$ will have a period of $2\pi/\omega$ and therefore can be written as a Fourier series expansion. For this purpose a Fourier series with complex Fourier coefficients is used:

$$P_n(t) = \sum_{k=-\infty}^{\infty} c_{k,n} e^{jk\omega t}, \qquad n = 0, 1, 2, \cdots \ . \tag{7.10}$$

It is obvious that probabilities cannot have complex values which implies that

$$c_{k,n} = c^*_{-k,n}, \qquad k = 0, 1, 2, \cdots \tag{7.11}$$

and that $c_{0,n}$ is real. Furthermore, the following probability constraints must be observed:

$$0 \leq P_n(t) \leq 1 \quad \text{for all } t \tag{7.12}$$

$$\sum_{n=0}^{\infty} P_n(t) = 1 \ . \tag{7.13}$$

Equation (7.13) can only hold at all times t if

$$\sum_{n=0}^{\infty} c_{k,n} = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.14}$$

Substituting equations (7.10) and (7.7) into equation (7.9) gives:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} &= -(\alpha+\beta\sin\omega t+\mu_n)\sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t} \\ &\quad +(\alpha+\beta\sin\omega t)\sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t} \\ &\quad +\mu_{n+1}\sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t} \quad . \end{aligned} \tag{7.15}$$

Equation (7.15) may be rewritten by writing the $\sin\omega t$ terms in exponential form:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} &= -(\alpha+\beta\frac{e^{j\omega t}-e^{-j\omega t}}{2j}+\mu_n)\sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t} \\ &\quad +(\alpha+\beta\frac{e^{j\omega t}-e^{-j\omega t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t} \\ &\quad +\mu_{n+1}\sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t}\end{aligned}$$

which implies that

$$\begin{aligned}jk\omega c_{k,n} &= -(\alpha+\mu_n)c_{k,n}-\frac{\beta}{2j}c_{k-1,n}+\frac{\beta}{2j}c_{k+1,n}+\alpha c_{k,n-1}+ \\ &\quad \frac{\beta}{2j}c_{k-1,n-1}-\frac{\beta}{2j}c_{k+1,n-1}+\mu_{n+1}c_{k,n+1} \quad .\end{aligned}$$

This is a non-linear complex recurrence relation for $c_{k,n}$ and can be rearranged as

$$\begin{aligned}(\alpha+\mu_n+jk\omega)c_{k,n} &= -\frac{\beta}{2j}c_{k-1,n}+\frac{\beta}{2j}c_{k+1,n}+\alpha c_{k,n-1}+ \\ &\quad \frac{\beta}{2j}c_{k-1,n-1}-\frac{\beta}{2j}c_{k+1,n-1}+\mu_{n+1}c_{k,n+1}\end{aligned}$$

or

$$(\alpha+\mu_n+jk\omega)c_{k,n} = \alpha c_{k,n-1}+j\frac{\beta}{2}(c_{k-1,n}-c_{k+1,n}-c_{k-1,n-1}+c_{k+1,n-1})+\mu_{n+1}c_{k,n+1} \quad . \tag{7.16}$$

Note that the recurrence relationship for any of the coefficients of the Fourier series involves 6 of its "neighbouring" coefficients. Equation (7.16) must be solved

subject to constraints given in equations (7.11), (7.12), (7.13) & (7.14). Note that the non-linearity of equation (7.16) makes it very difficult to find an analytical solution to the problem. A numerical solution however may be possible.

7.4.1 A Numerical Solution

To make a pictorial representation of equation (7.16) let us arrange the $c_{k,n}$ values in a rectangular array as shown in Figure 7.1.

	$n \rightarrow$						
	0	1	$\cdots$	$n-1$	n	$n+1$	$\cdots$
$\vdots$							
$k+1$				$c_{k+1,n-1}$	$c_{k+1,n}$	$c_{k+1,n+1}$	
k				$c_{k,n-1}$	$c_{k,n}$	$c_{k,n+1}$	
$k-1$				$c_{k-1,n-1}$	$c_{k-1,n}$	$c_{k-1,n+1}$	
$\vdots$							

Figure 7.1: Array of Fourier coefficients $c_{k,n}$

This array is infinite in the n dimension on one side and is doubly infinite in the k dimension. In order to solve this problem numerically, the array of Fourier series coefficients must be truncated to a finite array. As $c_{k,n} = c^*_{-k,n}$, one need only allocate space for the positive values of k (0 included) and use the following rules where applicable:

$$\mu_n = \begin{cases} \mu, & n > 0 \\ 0, & n = 0 \end{cases}$$

$$c_{k,n-1} = 0 \qquad \text{if } n = 0$$

$$c_{k-1,n} - c_{k+1,n} = \begin{cases} -2Im\{c_{k+1,n}\} & \text{if } k = 0 \\ c_{k-1,n} - c_{k+1,n} & \text{otherwise} \end{cases}$$

$$c_{k+1,n-1} - c_{k-1,n-1} = \begin{cases} 0 & \text{if } n = 0 \text{ (regardless of } k) \\ 2Im\{c_{k+1,n-1}\} & \text{if } k = 0 \text{ and } n \neq 0 \\ c_{k+1,n-1} - c_{k-1,n-1} & \text{otherwise .} \end{cases}$$

The range of values of n are truncated from $0 \cdots \infty$ to $0 \cdots m$. Also, for each value of n, the complex Fourier series for $P_n(t)$ is truncated so that k holds values between $-m$ to m. As initial values, the row corresponding to $k = 0$ is filled with the results for an $M/M/1$ system without the periodic component in the arrival rate, i.e. $\lambda(t) = \alpha$. The rest of the entries are initially set to zero. Equation (7.16) can then be applied recursively to improve the accuracy of the entries of the array.

The convergence of the Fourier series coefficients was found to be dependent on how the recurrence relationship is applied to the entries of the array. Initially the recursion was applied to the entries of the array on a row to row basis. All attempts failed and the results never converged. Next, recurring on the entries of the array on a block by block basis was examined. This time the results converged very quickly. It was found that under-relaxation or over-relaxation did not play a significant role in the speed of convergence.

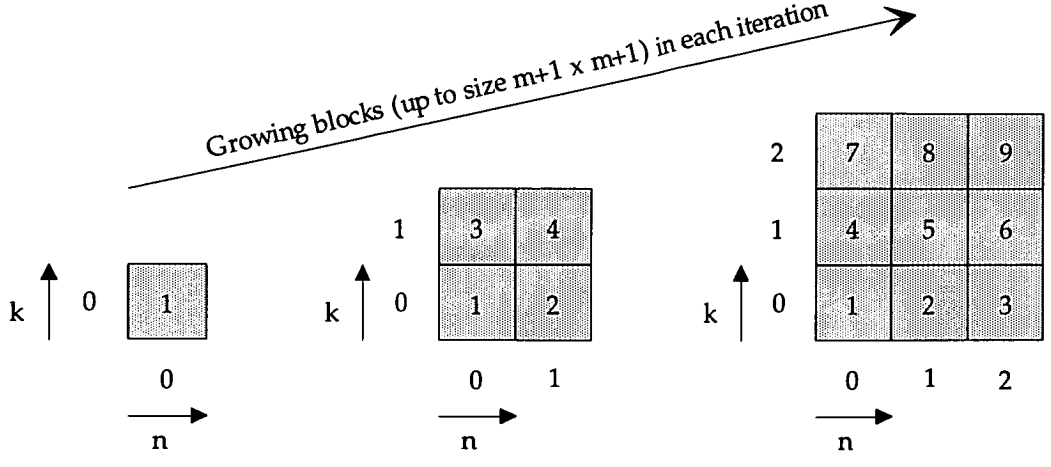


Figure 7.2: An illustration of the recurrence process in each iteration

The process of iterating on the entries of the array of Fourier series coefficients

is shown in Figure 7.2. Let us describe this process by assuming that the array has dimensions of 100 × 100. Let us define a counter called the *iteration index* and initialize it to zero. We start with a block size of 1 × 1 elements at the origin (i.e. $k = 0$ and $n = 0$) and the entry of this block is updated. Then, both dimensions of the block are increased by 1 and all the entries of the new block are updated. The order of updating the entries of a block that has more than one element is shown in Figure 7.2. The process of increasing the size of the block is continued, and all the entries of the new blocks are updated, until the block reaches its maximum size that is 100 × 100. The iteration index is then incremented by 1 and the whole process is repeated from the beginning. Therefore, the number of iterations quoted in this chapter and in the next chapter actually refer to the iteration index. For example, if it is said that the results converged after 50 iterations, it really means that the above process has been repeated until the iteration index reached 50.

After all the Fourier series coefficients have converged the probabilities of different system populations can be calculated from the following equation which is the truncated version of equation (7.10):

$$P_n(t) \simeq \sum_{k=-m}^{m} c_{k,n} e^{jkwt} \quad .$$

Now let $a_{k,n} + jb_{k,n}$ represent the complex coefficient $c_{k,n}$. Hence

$$\begin{aligned} P_n(t) &\simeq \sum_{k=-m}^{m} (a_{k,n} + jb_{k,n}) e^{jkwt} \\ &= a_{0n} + 2\sum_{k=1}^{m} a_{k,n} cos(kwt) - 2\sum_{k=1}^{m} b_{k,n} sin(kwt) \quad . \end{aligned} \tag{7.17}$$

The following parameters were selected to examine the numerical solution:

$$\begin{aligned} \alpha &= 1.0 \\ \beta &= 0.75 \\ \mu &= 2.0 \\ \omega &= 2\pi \end{aligned}$$

The dimensions of the array of Fourier series coefficients were truncated to 100 by 100 (i.e. $m = 99$). For different numbers of iterations, the results were compared and it was found that after about 100 iterations the Fourier series coefficients had converged very well and that iterating beyond 100 did not change the results significantly. In fact the convergence rate is so high that even after 10 iterations the results are very close to those obtained after 100 iterations. With 100 iterations, the probabilities of having different system populations are shown in Figures 7.3 to 7.6. Note that with $\omega = 2\pi$, 1 second corresponds to one complete cycle. As these figures show, the stationary probabilities of being in various states are periodic and have the same period as the arrival rate. Therefore it is appropriate for these probabilities to be called *Cyclo-Stationary Probabilities.* Also, what is normally referred to as mean system population will now have a cyclic nature and should therefore be called the *Cyclo-Stationary System Population.* This quantity is shown in Figure 7.7.

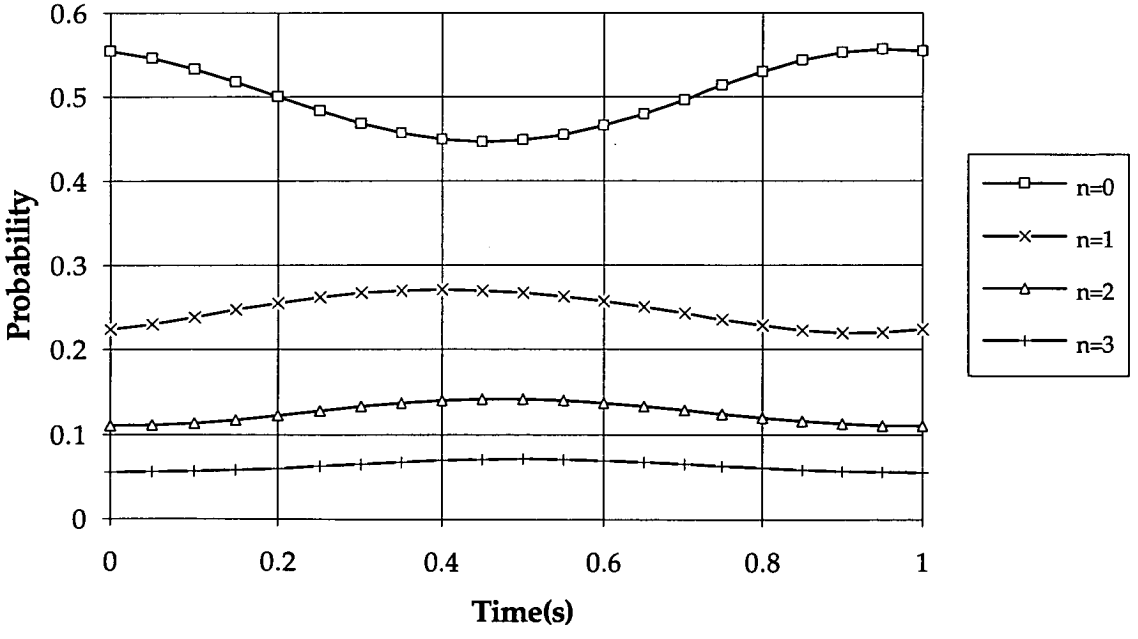


Figure 7.3: $P_n(t)$ *for* $n = 0 \cdots 3$ *with* $\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$

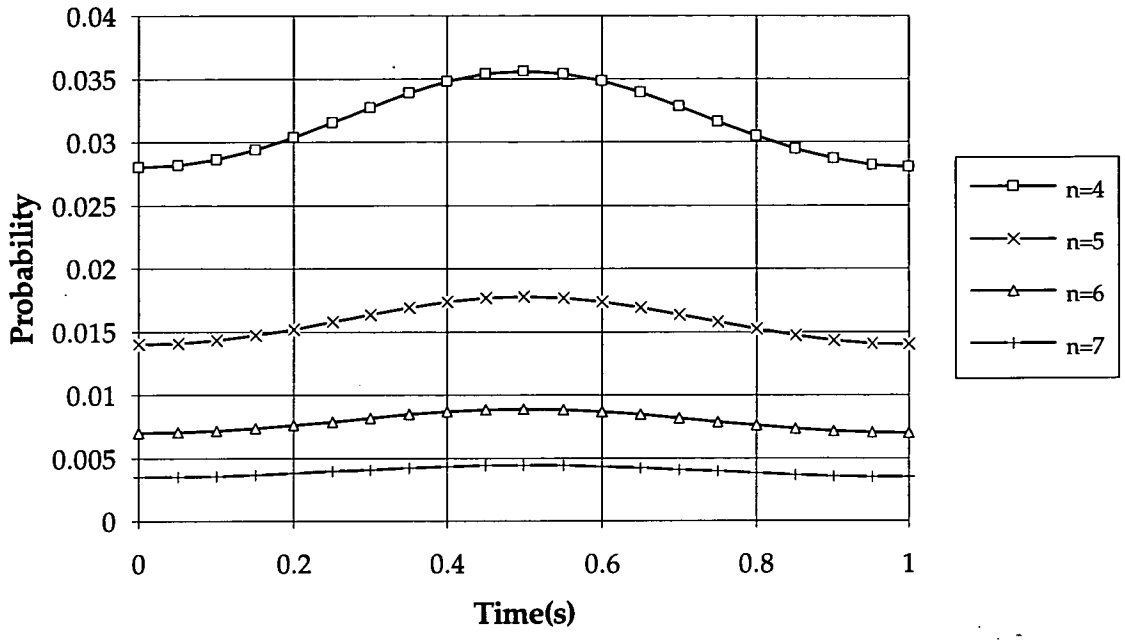


Figure 7.4: $P_n(t)$ *for* $n = 4 \cdots 7$ *with* $\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$

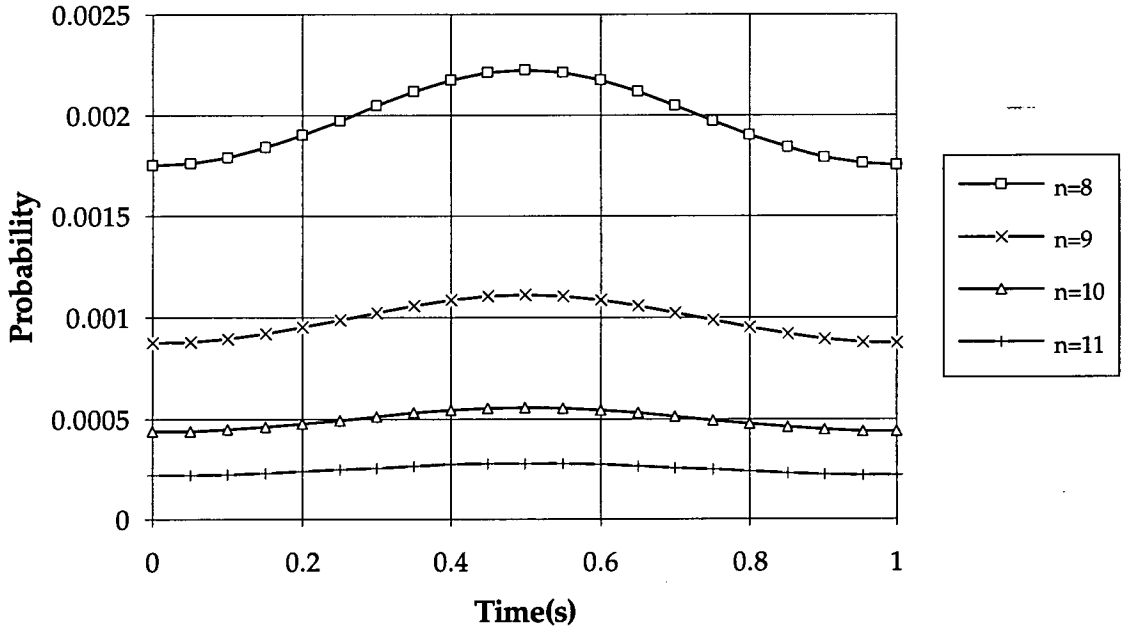


Figure 7.5: $P_n(t)$ *for* $n = 8 \cdots 11$ *with* $\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$

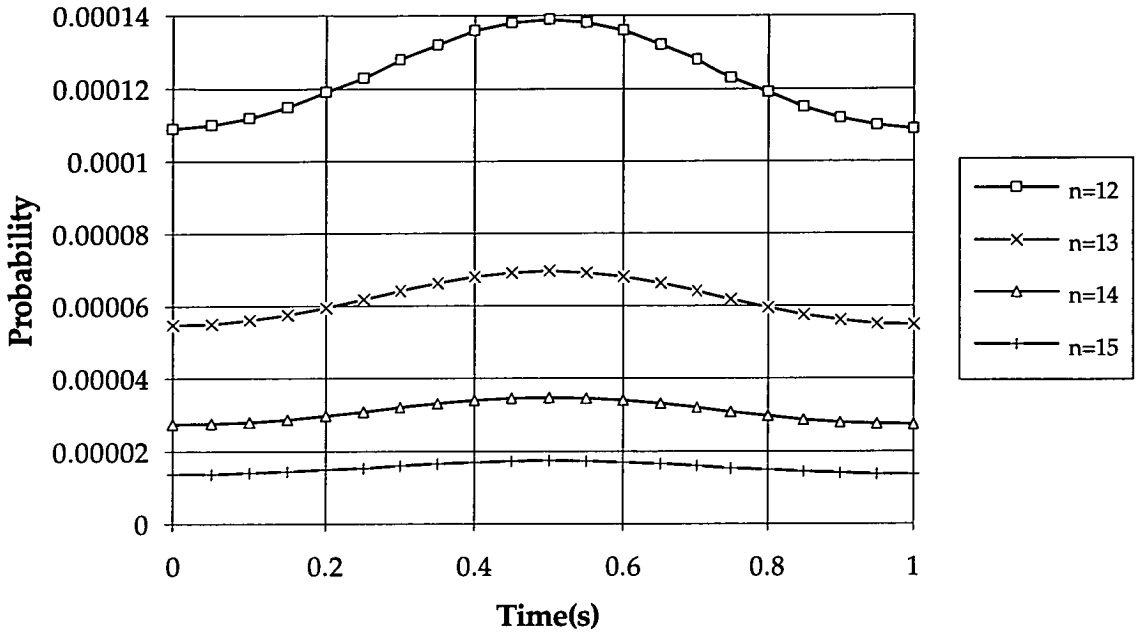


Figure 7.6: $P_n(t)$ for $n = 12 \cdots 15$ with $\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$

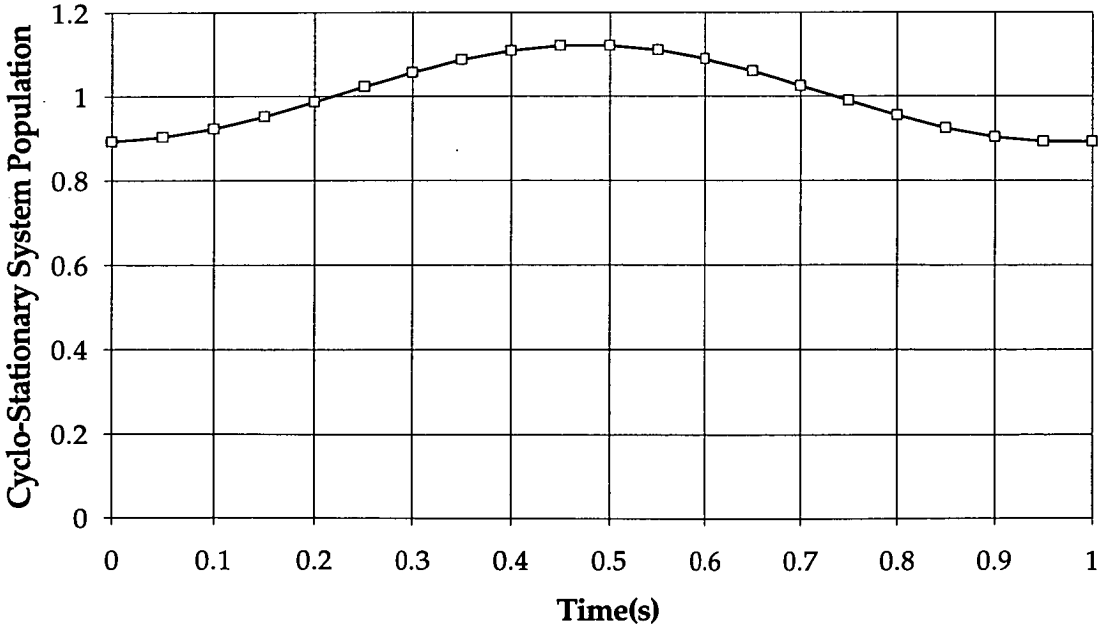


Figure 7.7: Cyclo-Stationary System Population with $\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$

7.4.2 Effect of the Frequency

With any given set of values for α, β and μ, the cyclo-stationary system population (and therefore other parameters such as typical peak durations) will depend on the frequency of variation of the arrival rate. For example, with lower frequencies the arrival rate spends a longer duration around each extreme in each cycle, hence giving more time to the system to react to the variations in the arrival rate. Therefore it is expected that the variation in cyclo-stationary system population will decrease as the frequency of variation of the arrival rate increases. Figure 7.8 confirms this argument.

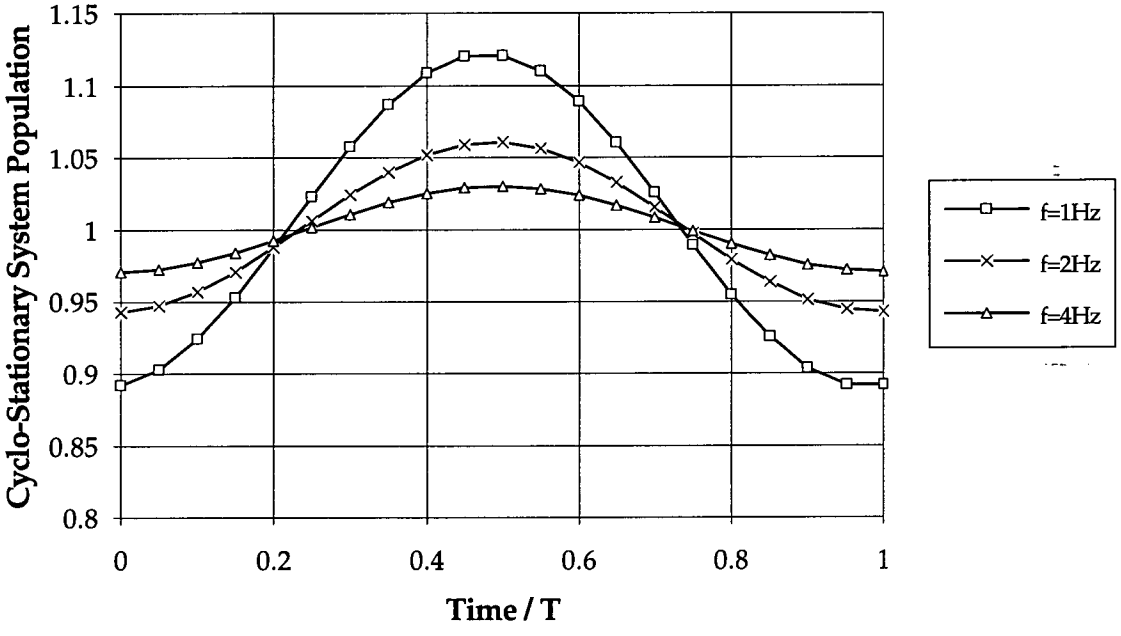


Figure 7.8: Cyclo-Stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$

7.4.3 Effects of Truncation

In this section results indicative of the accuracy of the numerical solution are presented. An obvious source of error in this numerical solution is due to the truncation of the array of Fourier series coefficients. To examine this error, the calculations were repeated with an array that was 5 times smaller in each dimension, i.e. 20 by 20 elements (or $m = 19$).

Figures 7.9 to 7.12 show a graphical comparison of the cyclo-stationary probabilities computed from the two different array sizes. Note that for values of n up to 10, both cases give almost identical results, but for higher values of n, the discrepancy caused by truncation becomes obvious. However, as Figure 7.13 shows, the cyclo-stationary system population is much less sensitive to the truncation errors. This is because with the given set of parameters lower queue sizes have much higher probabilities.

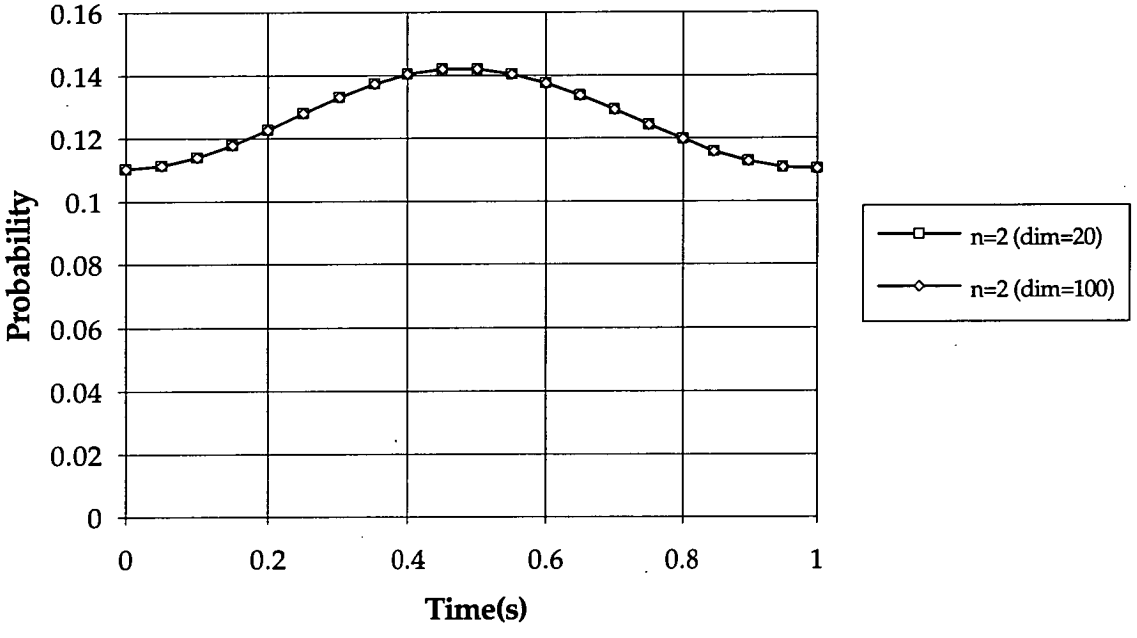


Figure 7.9: $P_2(t)$ for array dimensions of 20 and 100 ($\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$)

The discrepancy in probabilities due to truncation was examined with other parameters (and different values of m). The individual results are not shown here, but the conclusion was that in order to select an appropriate size for the array of Fourier series coefficients, the smallest probabilities of interest must be estimated, say P_1. Then the M/M/1 queueing results must be used to find the smallest value of n, say n_1, which conforms to $P(n_1) \leq P_1$. The dimension of the array of Fourier series coefficients must be several times larger than n_1 and should be determined after a few trials.

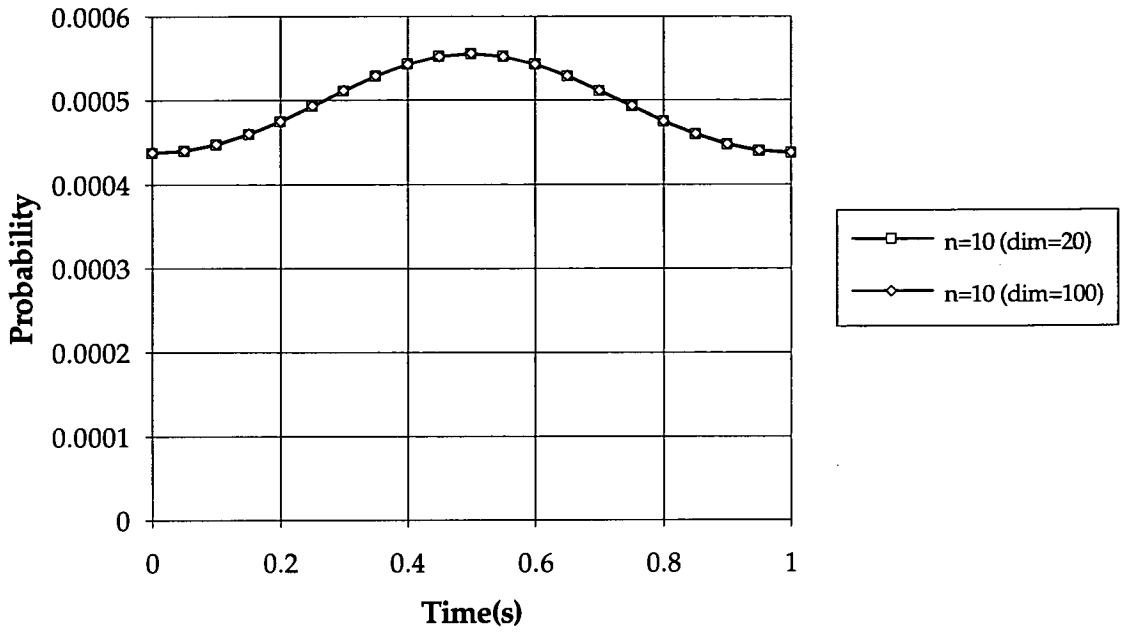


Figure 7.10: $P_{10}(t)$ *for array dimensions of 20 and 100* ($\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$)

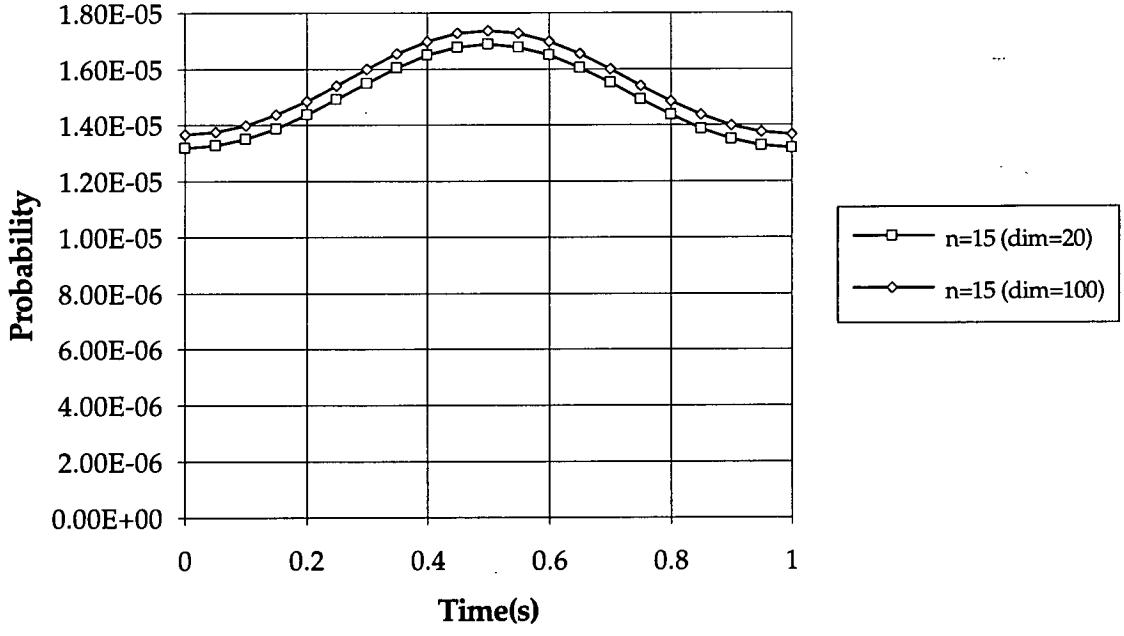


Figure 7.11: $P_{15}(t)$ *as a function of time for array dimensions of 20 and 100* ($\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$)

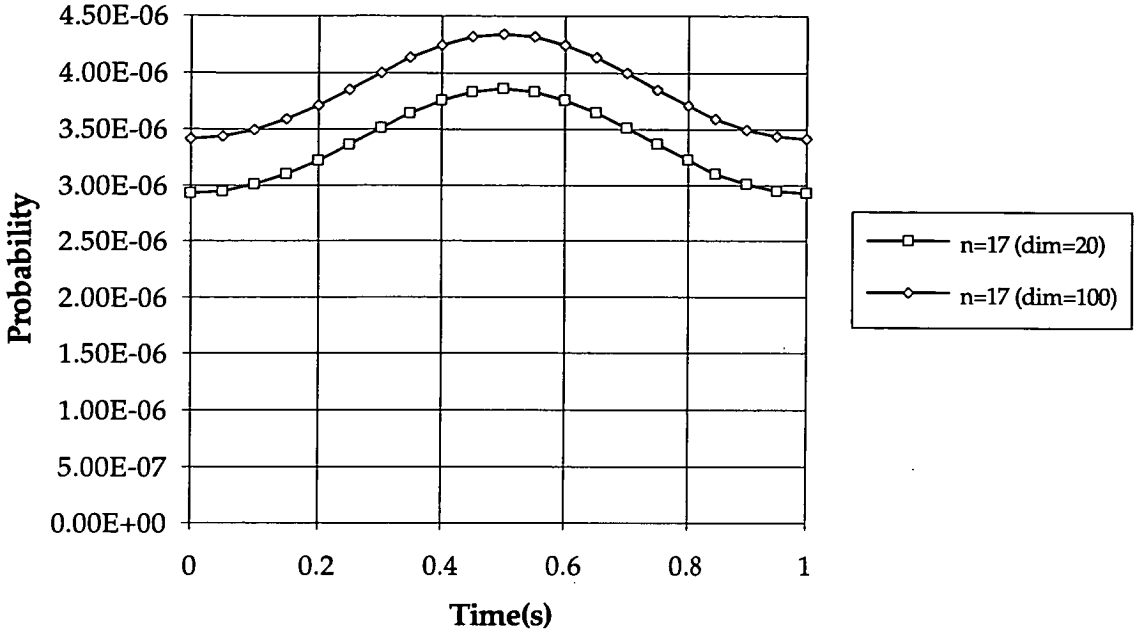


Figure 7.12: $P_{17}(t)$ for array dimensions of 20 and 100 ($\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$)

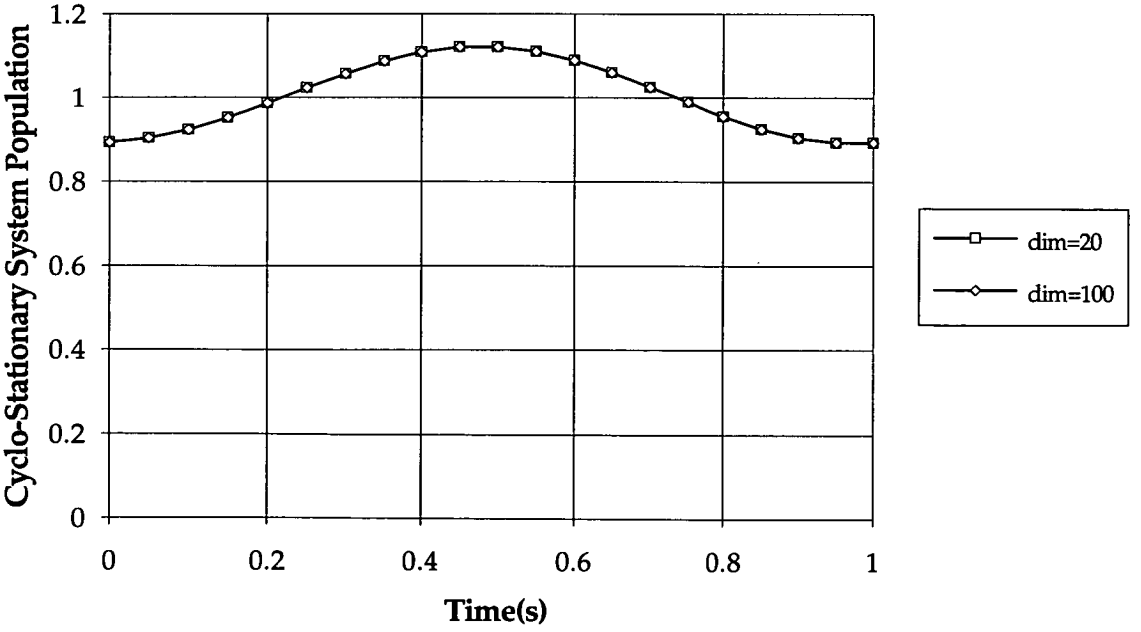


Figure 7.13: Cyclo-Stationary system population for array dimensions of 20 and 100 ($\alpha = 1.0$, $\beta = 0.75$, $\mu = 2.0$, $\omega = 2\pi$)

7.5 Generalised Periodic Arrivals

In this section the arrival rate is assumed to vary periodically but it has an arbitrary shape in each period rather than being restricted to a sinusoidal waveform. The arrival rate itself can therefore be written as a Fourier series:

$$\lambda(t) = \sum_{i=-\infty}^{\infty} \beta_i e^{ji\omega t} \tag{7.18}$$

where

$$\beta_i = \frac{1}{T} \int_{-T/2}^{T/2} \lambda(t) e^{-ji\omega t} dt \ . \tag{7.19}$$

Substituting equation (7.18) in (7.9) gives:

$$\begin{aligned} \sum_{k=-\infty}^{\infty} jk\omega c_{k,n} e^{jk\omega t} \ = \ & -\Big(\sum_{i=-\infty}^{\infty} \beta_i e^{ji\omega t} + \mu_n\Big) \sum_{k=-\infty}^{\infty} c_{k,n} e^{jk\omega t} \\ & + \sum_{i=-\infty}^{\infty} \beta_i e^{ji\omega t} \sum_{k=-\infty}^{\infty} c_{k,n-1} e^{jk\omega t} \\ & + \mu_{n+1} \sum_{k=-\infty}^{\infty} c_{k,n+1} e^{jk\omega t} \ . \end{aligned}$$

Therefore

$$jk\omega c_{k,n} = -\sum_{i=-\infty}^{\infty} \beta_i c_{k-i,n} - \mu_n c_{k,n} + \sum_{i=-\infty}^{\infty} \beta_i c_{k-i,n-1} + \mu_{n+1} c_{k,n+1}$$

or

$$(\mu_n + jk\omega) c_{k,n} = -\sum_{i=-\infty}^{\infty} \beta_i c_{k-i,n} + \sum_{i=-\infty}^{\infty} \beta_i c_{k-i,n-1} + \mu_{n+1} c_{k,n+1} \ . \tag{7.20}$$

This is a recurrence relationship where each entry in the array depends on an infinite number of other entries in that array. However, if the Fourier series describing the periodic arrival rate is limited to a finite number of harmonics then the recursive relationship will have a finite number of terms. The truncated version of equation (7.20) is:

$$(\mu_n + jk\omega) c_{k,n} = -\sum_{i=-l}^{l} \beta_i c_{k-i,n} + \sum_{i=-l}^{l} \beta_i c_{k-i,n-1} + \mu_{n+1} c_{k,n+1} \tag{7.21}$$

where l is the limit on the number of harmonics.

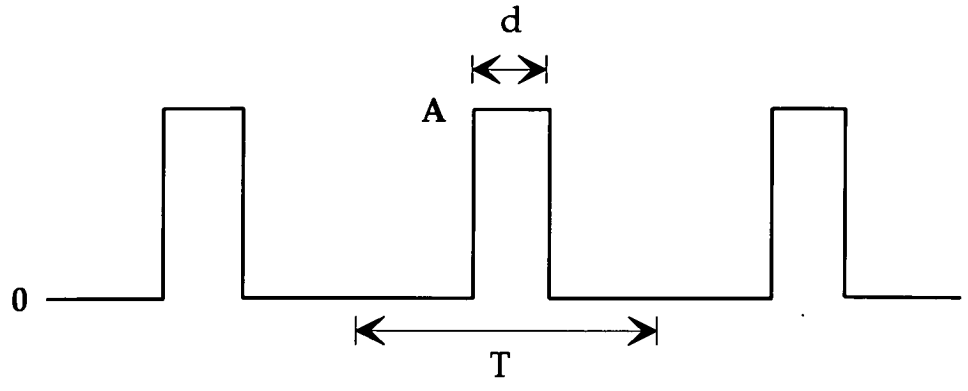


Figure 7.14: Square waveform cyclo-stationary arrival rate

7.5.1 Example: Square Waveform

Let us consider the case where the mean arrival rate has the shape of a square waveform as shown in Figure 7.14 with a period of T seconds, pulse width of d seconds, minimum of zero and maximum of A. The coefficients of the Fourier series for this arrival rate are calculated as:

$$\begin{aligned}\beta_i &= \frac{1}{T}\int_{-T/2}^{T/2}\lambda(t)e^{-ji\omega t}dt \\ &= \frac{1}{T}\int_{-d/2}^{d/2}Ae^{-ji\omega t}dt \\ &= \frac{Ad}{T}\frac{\sin(i\omega d/2)}{i\omega d/2}\end{aligned}$$

The following parameters were selected to examine the numerical solution:

$$\begin{aligned}A &= 1.0 \\ d &= 0.2 \\ \mu &= 1.2 \\ \omega &= 2\pi \\ l &= 40 \\ m &= 99\end{aligned}$$

With the first 40 harmonics of the arrival rate waveform, the approximated ar-

rival rate is shown in Figure 7.15. The overshoots and undershoots seen in Figure 7.15 is due to *Gibbs' phenomenon* [141]. This phenomenon implies that because the arrival rate is discontinuous at $t = -0.1$ and at $t = 0.1$, then in the vicinity of these points the Fourier Series approximation oscillates regardless of the number of Fourier Series terms used. Nevertheless, the duration of oscillation decreases as the number of Fourier series terms increases. This phenomenon however does not seem to have a visible effect on the results presented later in this section.

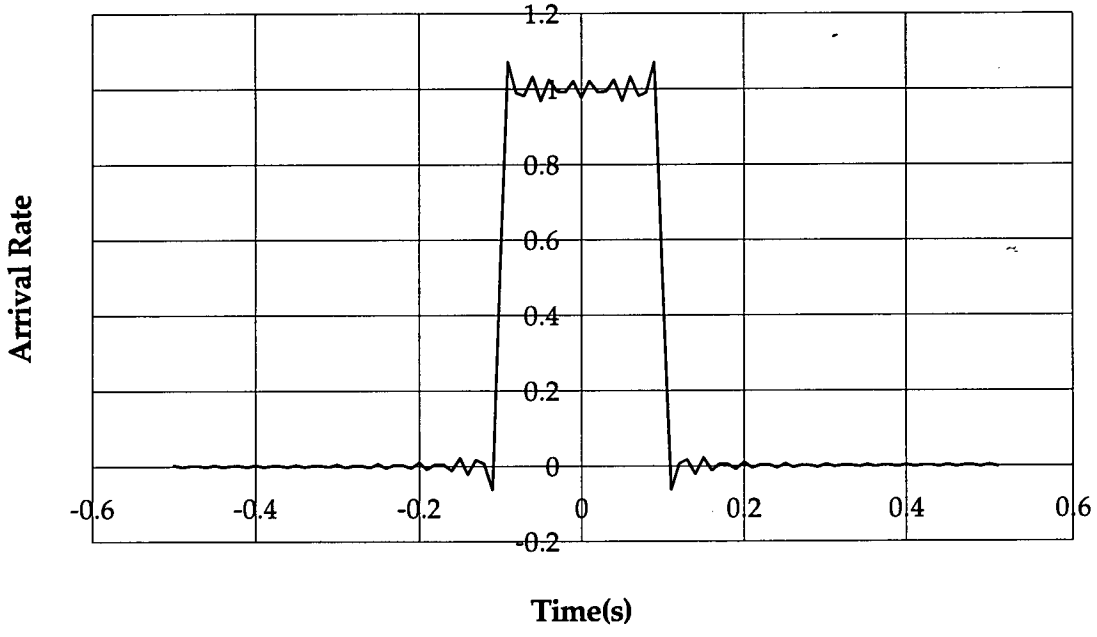


Figure 7.15: Cyclo-stationary input rate approximated with the first 40 harmonics

Although the computation time for the given example increases markedly compared to the case where the arrival rate has a fixed term and a single *sine* term, the actual convergence is still quite good. Using double precision floats on an IBM RISC 6000 machine, the numerical solution could calculate probabilities as small as 1×10^{-20}. Let us first show that the results presented here have converged. Figures 7.16 & 7.17 show the cyclo-stationary probability of a system population of 15. The results of these figures have been obtained after 14 iterations and 262 iterations respectively. Note that there are sub-iterations within each iteration. Therefore the number of iterations quoted here is only a comparative index rather. It is not the actual number of times that equation (7.21) has been applied

to the entries of the array of Fourier series coefficients.

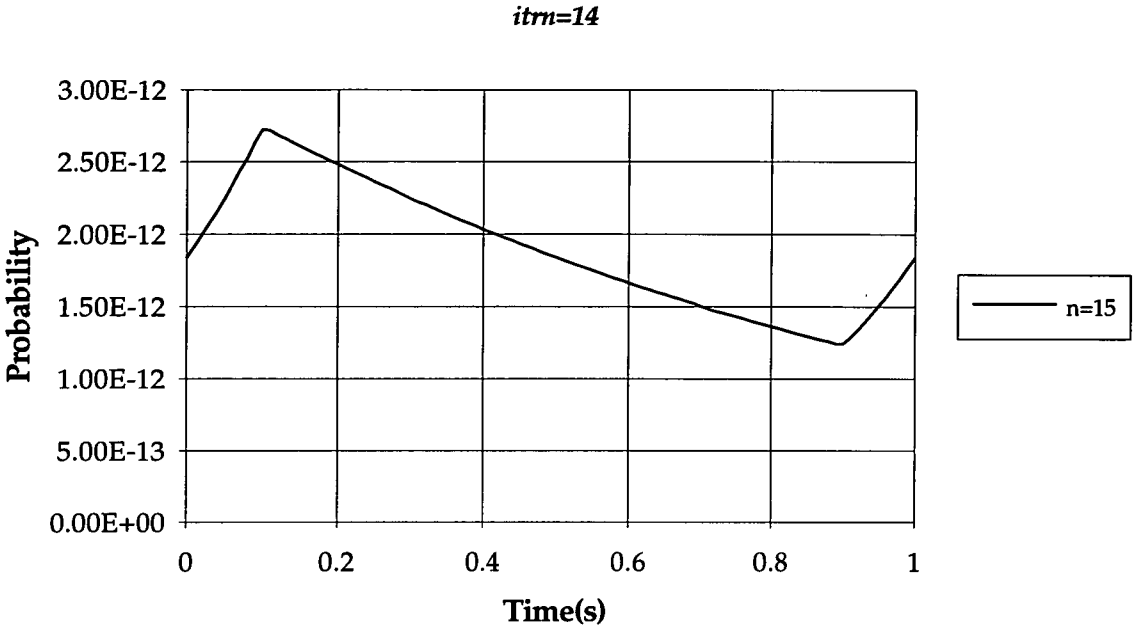


Figure 7.16: $P_{15}(t)$ with the iteration index = 14

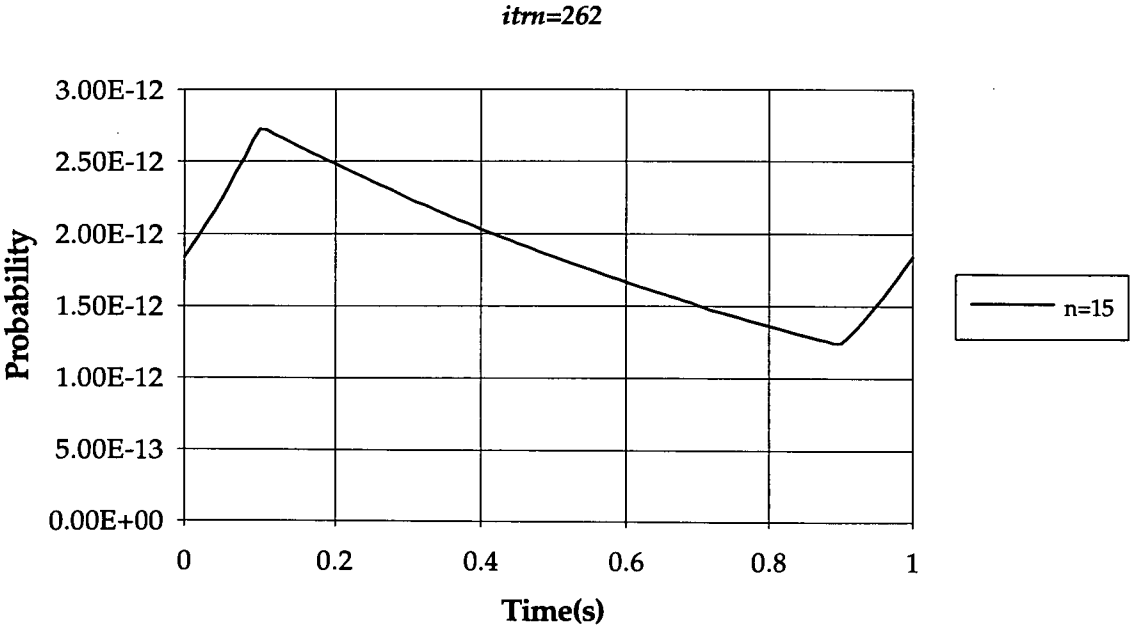


Figure 7.17: $P_{15}(t)$ with the iteration index = 262

The reason for showing the results of Figures 7.16 & 7.17 on two different graphs

is that the numbers are almost identical and with a single graph the two lines cannot be differentiated.

It is logical to assume that the probability of zero system population must decrease during the pulse of the arrival waveform and that it must increase during the space (zero arrival rate) of the arrival waveform. All other probabilities could be expected to behave in the opposite way, i.e. for $n \neq 0$, $P_n(t)$ should increase during the pulse and decrease during the space of the arrival waveform. These assumptions are validated by Figures 7.18 and 7.19 which show the cyclo-stationary probabilities for system populations of 0, 1, 2 and 3. Further results for cyclo-stationary system population is shown in Figure 7.20.

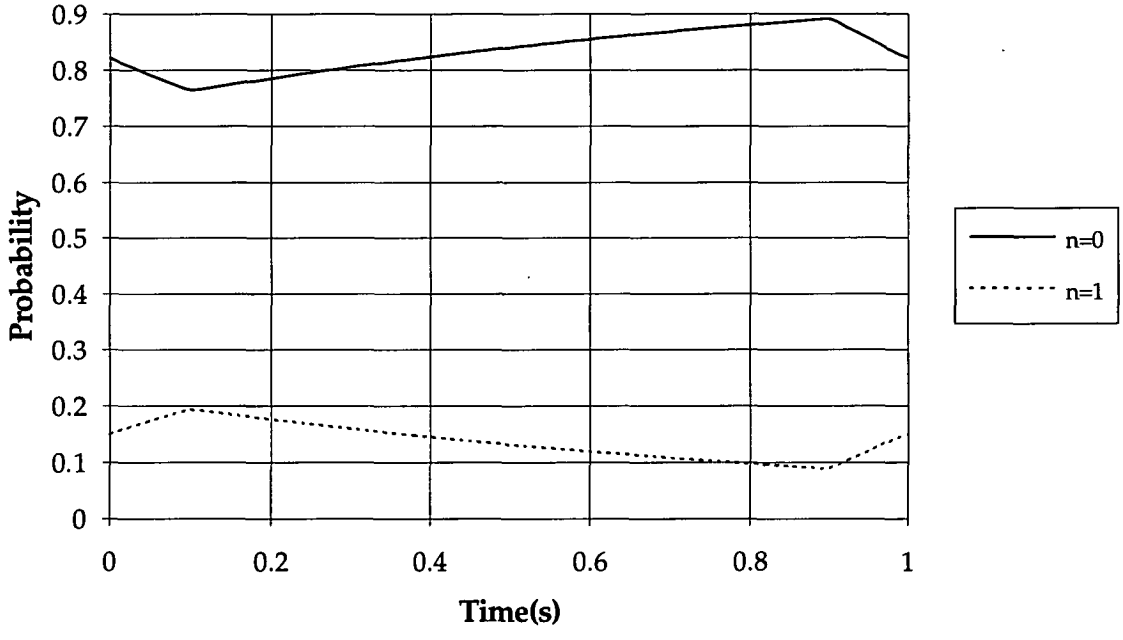


Figure 7.18: $P_n(t)$ $(n = 0, 1)$ for the square waveform arrival rate

7.6 Summary

In this chapter a method of analysis has been presented for queueing analysis of the systems that have periodic elements in the arrival rate of their traffic. This technique is particularly useful when the period of the cyclic traffic is not vastly greater than the interarrival times. A numerical solution was initially developed

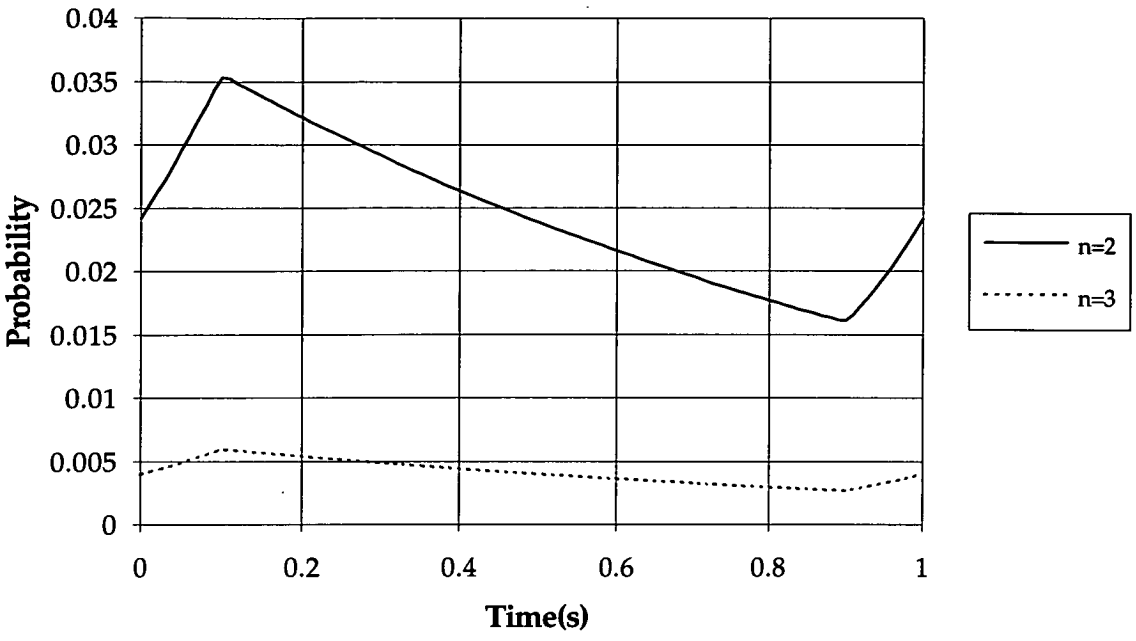


Figure 7.19: $P_n(t)$ $(n = 2, 3)$ for the square waveform arrival rate

for cyclo-stationary arrivals with a mean arrival rate that had the shape of a sinusoid. The method of analysis was based on using Fourier series (with complex coefficients) to describe the cyclo-stationary probabilities for different states of the system. System performance measures can be calculated from these coefficients. The analysis resulted in a non-linear, complex recurrence relationship for each Fourier series coefficient in terms of its "neighbouring" coefficients. We found that the convergence of the Fourier series coefficients was dependent on how the recurrence relationship was applied to the array of coefficients. A particular method based on recursive estimation of the coefficients on a block by block basis resulted in very fast convergence.

The effect of the arrival rate frequency on the performance of the queue was considered. With all other parameters fixed, it was found that the variation in the cyclo-stationary system population decreased as the frequency of variation of the arrival rate increases.

The effect of truncating the Fourier series coefficients on the accuracy of the results was considered. It was suggested that in order to select an appropriate size

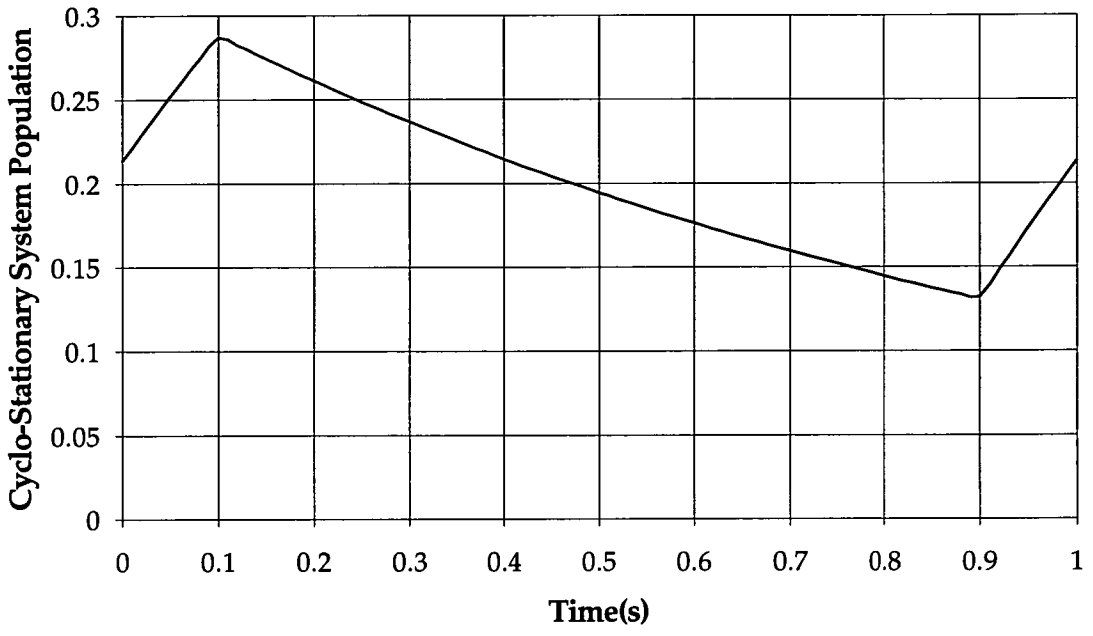


Figure 7.20: Cyclo-stationary system population for the square waveform arrival rate

for the array of Fourier series coefficients, the smallest probabilities of interest must be determined, say P_1. The M/M/1 queueing results must then be used to find the smallest system population, say n_1, such that $P(n_1) \leq P_1$. The dimensions of the array of Fourier series coefficients should be several times larger than n_1 and should be determined after a few trials.

The method of analysis was extended to cater for arbitrary shapes of cyclo-stationary arrival rates. The arrival rate itself was described in the form of a truncated Fourier series with $2l + 1$ coefficients. The analysis resulted in a recurrence relationship for Fourier series coefficient $c_{k,n}$ in terms of $4l + 2$ other coefficients. The computation time increased considerably, but the convergence was still satisfactory. An example was considered where the cyclo-stationary arrival rate had the shape of a square waveform and performance results were generated.

Chapter 8

Queues with Periodic Arrival & Service Rates

8.1 Introduction

In the previous chapter, a method of analysis was presented for the study of the queueing systems that have random arrivals with periodically varying mean arrival rate. It was shown how Fourier series can be used as a tool to analyse such a system. In this chapter a more general scenario of periodicities in a queueing system is presented. Here, both the mean arrival rate and mean service rate vary periodically. An example of where periodic mean service rate is applicable is cyclic suspension of service while other traffic is handled. Such service strategies are commonly found in telecommunication and computer networks. Initially we consider the case where the cyclo-stationary arrival rate and the cyclo-stationary service rate are sinusoidal. A numerical solution is then presented for identical arrival rate and service rate frequencies. Next, arbitrary frequencies are assumed for the arrival rate and the service rates and the necessary conditions for achieving a numerical solution are studied. Finally, the analysis is then extended for generalised shapes of cyclo-stationary arrivals and cyclo-stationary service rates with arbitrary frequencies.

8.2 Sinusoidal Periodic Input & Periodic Output

As in the last chapter, random arrivals are assumed with a mean value that oscillates sinusoidally around a fixed value:

$$\lambda(t) = \alpha + \beta \sin \omega_1 t \ . \tag{8.1}$$

Furthermore, the service completion time is taken to be random, with the mean service rate being a time varying function given by

$$\mu(t) = \tau + \gamma \sin \omega_2 t \ . \tag{8.2}$$

Although it is assumed that $\mu(t)$ is independent of n, the n subscript is used to avoid the need for individual treatment of the $n = 0$ case. Thus

$$\mu_n(t) = \tau_n + \gamma_n \sin \omega_2 t \tag{8.3}$$

and

$$\tau_n = \begin{cases} \tau, & n > 0 \\ 0, & n = 0 \end{cases} \tag{8.4}$$

$$\gamma_n = \begin{cases} \gamma, & n > 0 \\ 0, & n = 0 \ . \end{cases} \tag{8.5}$$

With these provisions, the following differential equation is applicable:

$$\frac{dP_n(t)}{dt} = -\{\lambda(t) + \mu_n(t)\}P_n(t) + \lambda_{n-1}(t)P_{n-1}(t) + \mu_{n+1}(t)P_{n+1}(t) \ . \tag{8.6}$$

Given the above definitions and assuming a stable queue, under certain conditions that will be outlined later, the probabilities of being in various states will, in the long run, acquire periodicity. At this stage $2\pi/\omega$ is taken to be the period of these probabilities. Although the conditions under which the probabilities will be periodic are not specified, for the time being periodicity of the probabilities is assumed. Therefore each probability can be written as a Fourier series as before:

$$P_n(t) = \sum_{k=-\infty}^{\infty} c_{k,n} e^{jk\omega t}, \quad n = 0, 1, 2, \cdots \ . \tag{8.7}$$

The constraints given in equations (7.11) to (7.14) hold here as well and will not be repeated. Substituting equations (8.1), (8.3) and (8.7) into equation (8.6) gives:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} &= -(\alpha+\beta\sin\omega_1 t+\tau_n+\gamma_n\sin\omega_2 t)\sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t}\\ &\quad +(\alpha+\beta\sin\omega_1 t)\sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t}\\ &\quad +(\tau_{n+1}+\gamma_{n+1}\sin\omega_2 t)\sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t}\ . \end{aligned} \tag{8.8}$$

Writing sin terms in their exponential form gives:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} &= -(\alpha+\beta\frac{e^{j\omega_1 t}-e^{-j\omega_1 t}}{2j}+\tau_n+\gamma_n\frac{e^{j\omega_2 t}-e^{-j\omega_2 t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t}\\ &\quad +(\alpha+\beta\frac{e^{j\omega_1 t}-e^{-j\omega_1 t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t}\\ &\quad +(\tau_{n+1}+\gamma_{n+1}\frac{e^{j\omega_2 t}-e^{-j\omega_2 t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t}\ . \end{aligned} \tag{8.9}$$

At this stage, it is appropriate to study the cases for which the assumption of periodic probabilities is valid.

8.2.1 Identical Input & Output Frequencies

Let us assume that the frequencies of the cyclo-stationary arrival rate and the cyclo-stationary service rate are the same, i.e. $\omega_1 = \omega_2 = \omega$:

$$\lambda(t) = \alpha + \beta\sin\omega t \tag{8.10}$$

$$\mu_n(t) = \tau_n + \gamma_n\sin\omega t\ . \tag{8.11}$$

Hence equation (8.9) becomes:

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} &= -(\alpha+\beta\frac{e^{j\omega t}-e^{-j\omega t}}{2j}+\tau_n+\gamma_n\frac{e^{j\omega t}-e^{-j\omega t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t}\\ &\quad +(\alpha+\beta\frac{e^{j\omega t}-e^{-j\omega t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t}\\ &\quad +(\tau_{n+1}+\gamma_{n+1}\frac{e^{j\omega t}-e^{-j\omega t}}{2j})\sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t}\ . \end{aligned} \tag{8.12}$$

Therefore

$$\begin{aligned} jk\omega c_{k,n} &= -(\alpha+\tau_n)c_{k,n} - \frac{\beta+\gamma_n}{2j}c_{k-1,n} + \frac{\beta+\gamma_n}{2j}c_{k+1,n} \\ &\quad +\alpha c_{k,n-1} + \frac{\beta}{2j}c_{k-1,n-1} - \frac{\beta}{2j}c_{k+1,n-1} \\ &\quad +\tau_{n+1}c_{k,n+1} + \frac{\gamma_{n+1}}{2j}c_{k-1,n+1} - \frac{\gamma_{n+1}}{2j}c_{k+1,n+1} \end{aligned}$$

or

$$\begin{aligned} (\alpha+\tau_n+jk\omega)c_{k,n} &= \alpha c_{k,n-1} + \tau_{n+1}c_{k,n+1} + \frac{\beta+\gamma_n}{2j}(c_{k+1,n} - c_{k-1,n}) \\ &\quad +\frac{\beta}{2j}(c_{k-1,n-1} - c_{k+1,n-1}) + \frac{\gamma_{n+1}}{2j}(c_{k-1,n+1} - c_{k+1,n+1}) \end{aligned}$$

or

$$\begin{aligned} (\alpha+\tau_n+jk\omega)c_{k,n} &= \alpha c_{k,n-1} + \tau_{n+1}c_{k,n+1} \\ &\quad -j\frac{\beta}{2}(c_{k+1,n} - c_{k-1,n} + c_{k-1,n-1} - c_{k+1,n-1}) \\ &\quad -j\frac{\gamma_n}{2}(c_{k+1,n} - c_{k-1,n}) + j\frac{\gamma_{n+1}}{2}(c_{k+1,n+1} - c_{k-1,n+1}) \ . \end{aligned} \tag{8.13}$$

We note that the recurrence relationship for any of the coefficients of the Fourier Series involves 8 of its neighbouring coefficients as opposed to 6 neighbouring coefficients in equation (7.16). As an example let us take the following set of parameters to test this model:

$$\begin{aligned} \alpha &= 1.0 \\ \beta &= 0.75 \\ \tau &= 2.0 \\ \gamma &= 1.0 \\ \omega_2 &= \omega_1 = \omega = 2\pi \ . \end{aligned}$$

Although due to extra floating point calculations the computation time for this model was much longer compared to the case of periodic input only, the speed of convergence did not seem to have been affected. Figure 8.1 shows probabilities for system populations of 0, 1, 2 & 3 for the above parameters. The cyclo-stationary system population for this example is shown in Figure 8.2.

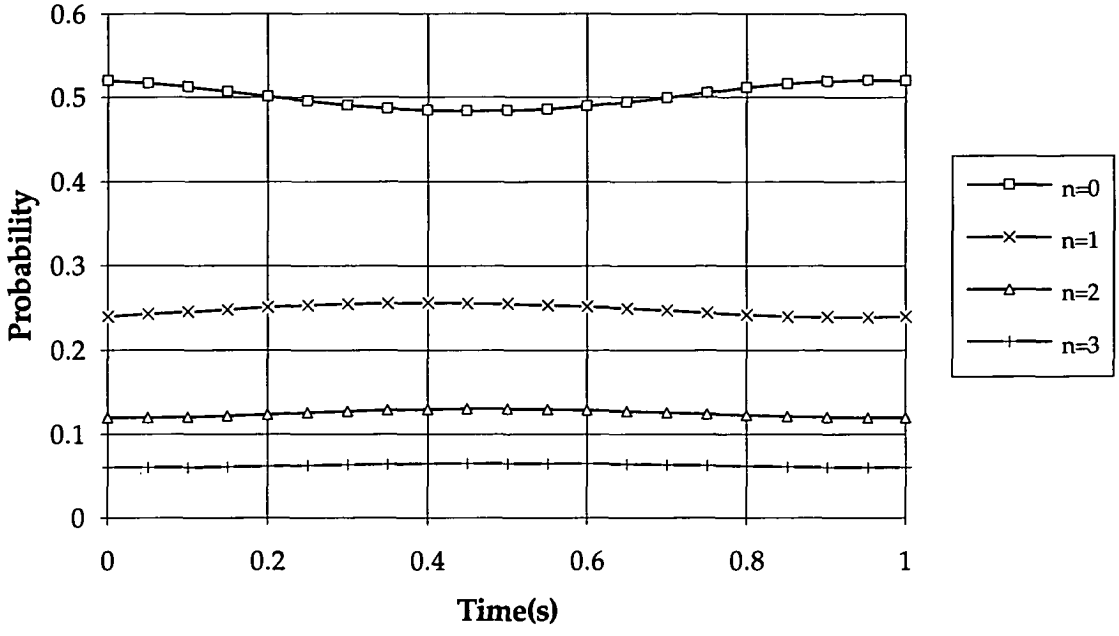


Figure 8.1: $P_n(t)$ *for* $n = 0 \cdots 3$ *with* $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $\omega_2 = \omega_1 = \omega = 2\pi$

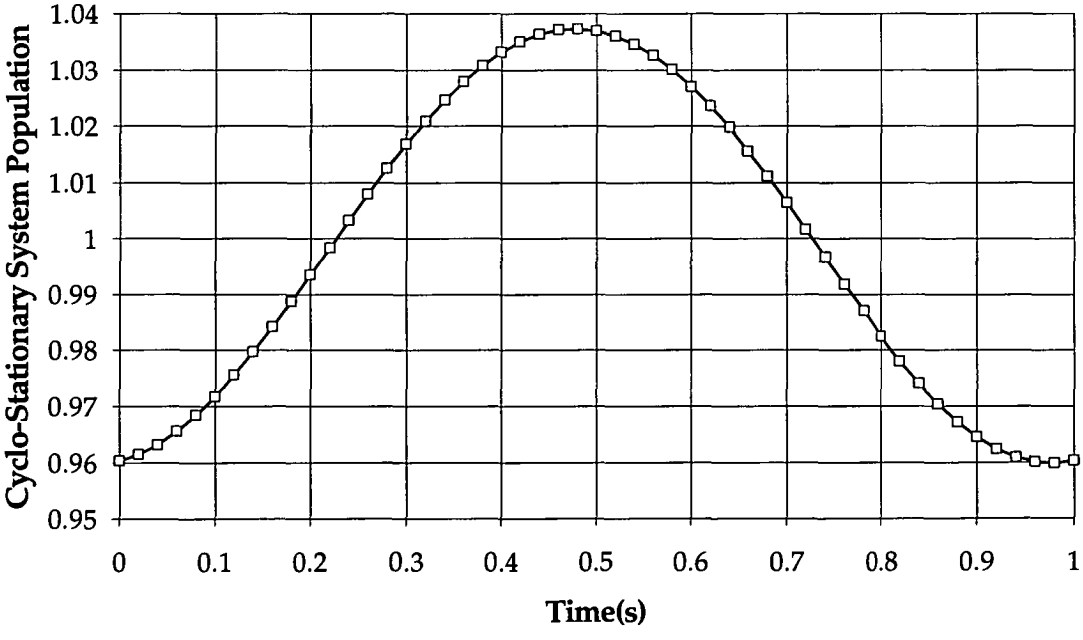


Figure 8.2: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $\omega_2 = \omega_1 = \omega$

8.2.2 Different Input & Output Frequencies

The most general condition under which the relationship given by equation (8.9) can yield a numerical solution occurs when both ω_1 and ω_2 are multiples of a frequency, say ω, with some phase shifts. Hence, the arrival rate and the service rate may be rewritten as follows.

$$\lambda(t) = \alpha + \beta \sin(a_1 \omega t + \phi_1) \tag{8.14}$$

$$\mu_n(t) = \tau_n + \gamma_n \sin(a_2 \omega t + \phi_2) \tag{8.15}$$

where a_1 and a_2 are integers and ϕ_1 and ϕ_2 are phase shifts in radians. With these conditions, equation (8.9) becomes:

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} jk\omega c_{k,n} e^{jk\omega t} &= -(\alpha + \beta \frac{e^{j(a_1\omega t+\phi_1)} - e^{-j(a_1\omega t+\phi_1)}}{2j} \\
&\quad + \tau_n + \gamma_n \frac{e^{j(a_2\omega t+\phi_2)} - e^{-j(a_2\omega t+\phi_2)}}{2j}) \sum_{k=-\infty}^{\infty} c_{k,n} e^{jk\omega t} \\
&\quad + (\alpha + \beta \frac{e^{j(a_1\omega t+\phi_1)} - e^{-j(a_1\omega t+\phi_1)}}{2j}) \sum_{k=-\infty}^{\infty} c_{k,n-1} e^{jk\omega t} \\
&\quad + (\tau_{n+1} + \gamma_{n+1} \frac{e^{j(a_2\omega t+\phi_2)} - e^{-j(a_2\omega t+\phi_2)}}{2j}) \sum_{k=-\infty}^{\infty} c_{k,n+1} e^{jk\omega t}
\end{aligned} \tag{8.16}$$

or

$$\begin{aligned}
jk\omega c_{k,n} &= -(\alpha + \tau_n) c_{k,n} - \frac{\beta e^{j\phi_1}}{2j} c_{k-a_1,n} + \frac{\beta e^{-j\phi_1}}{2j} c_{k+a_1,n} - \frac{\gamma_n e^{j\phi_2}}{2j} c_{k-a_2,n} \\
&\quad + \frac{\gamma_n e^{-j\phi_2}}{2j} c_{k+a_2,n} + \alpha c_{k,n-1} + \frac{\beta e^{j\phi_1}}{2j} c_{k-a_1,n-1} - \frac{\beta e^{-j\phi_1}}{2j} c_{k+a_1,n-1} \\
&\quad + \tau_{n+1} c_{k,n+1} + \frac{\gamma_{n+1} e^{j\phi_2}}{2j} c_{k-a_2,n+1} - \frac{\gamma_{n+1} e^{-j\phi_2}}{2j} c_{k+a_2,n+1}
\end{aligned}$$

or

$$\begin{aligned}
(\alpha + \tau_n + jk\omega) c_{k,n} &= \alpha c_{k,n-1} + \tau_{n+1} c_{k,n+1} + \frac{\beta e^{j\phi_1}}{2j} (c_{k-a_1,n-1} - c_{k-a_1,n}) \\
&\quad + \frac{\beta e^{-j\phi_1}}{2j} (c_{k+a_1,n} - c_{k+a_1,n-1}) - \frac{\gamma_n e^{j\phi_2}}{2j} c_{k-a_2,n} + \frac{\gamma_n e^{-j\phi_2}}{2j} c_{k+a_2,n} \\
&\quad + \frac{\gamma_{n+1} e^{j\phi_2}}{2j} c_{k-a_2,n+1} - \frac{\gamma_{n+1} e^{-j\phi_2}}{2j} c_{k+a_2,n+1} \; .
\end{aligned} \tag{8.17}$$

We note that the recurrence relationship for any of the coefficients of the Fourier Series involves 10 of the neighbouring coefficients as opposed to 8 neighbouring coefficients in equation (8.13). Equation (8.13) implies that the larger are the values of a_1 and a_2, the slower will be the speed of convergence. The reason is that with smaller values of a_1 and a_2, the neighbouring coefficients used in the calculation of $C_{k,n}$ have been updated more recently. The following set of parameters were selected to test the convergence of the new model:

$$\begin{aligned} \alpha &= 1.0 \\ \beta &= 0.75 \\ \tau &= 2.0 \\ \gamma &= 1.0 \\ \omega &= 2\pi \\ a_1 &= 2 \\ a_2 &= 3 \ . \end{aligned}$$

For different values of ϕ_1 and ϕ_2, equation (8.17) was applied recursively to find the Fourier series coefficients of the cyclo-stationary probabilities. Figure 8.3 shows results for cyclo-stationary probabilities of having system populations of $0 \cdots 3$ with $\phi_1 = \pi/4$ and $\phi_2 = -\pi/4$. It should be noted that the probability of zero system population behaves oppositely to the probabilities of non-zero system populations.

The cyclo-stationary system populations for different values of ϕ_1 and ϕ_2 are shown in Figures 8.4 to 8.10. These figures show that the magnitude and sense of the phase shifts of the input rate and the output rate of the queue can have a major effect on the performance of the system. If the frequencies of the arrival rate and the service rate were the same, i.e. $\omega_1 = \omega_2$, and if both ϕ_1 and ϕ_2 were increased or decreased by $\Delta\phi$, then all performance measures would have the same magnitude, but they would be shifted in time by $\frac{\Delta\phi}{2\pi}T$, where $T = 1/f$. However, this argument would no longer be valid if $\omega_1 \neq \omega_2$ in which case a phase shift of $\Delta\phi$ would correspond to different shifts in time for the arrival rate and the service rate.

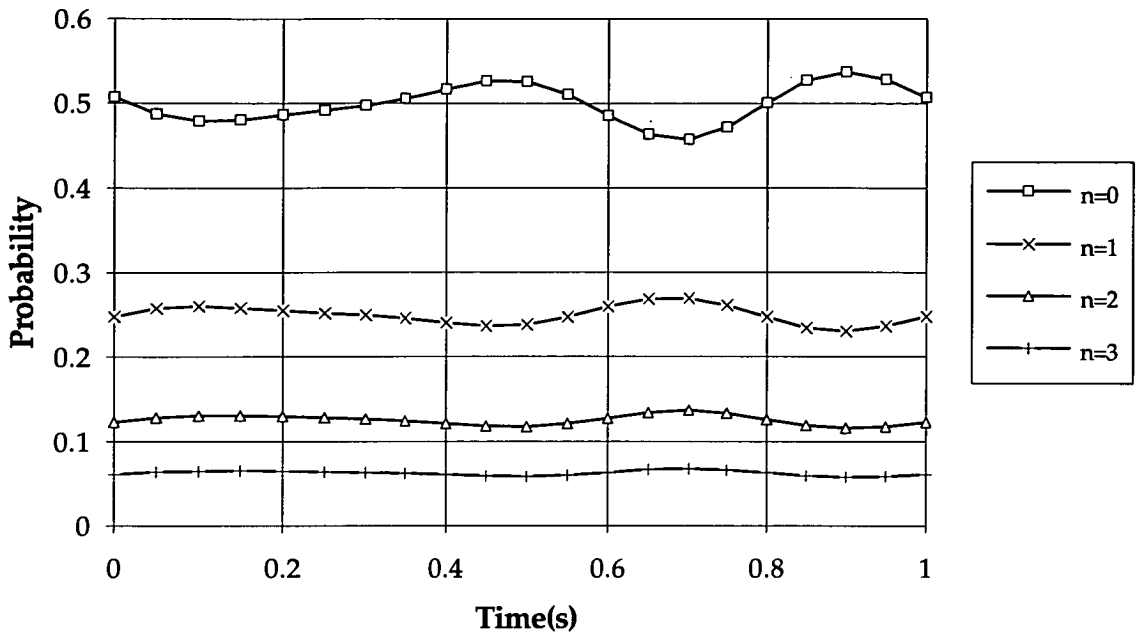


Figure 8.3: $P_n(t)$ *for* $n = 0 \cdots 3$ *with* $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = \pi/4$, $\phi_2 = -\pi/4$

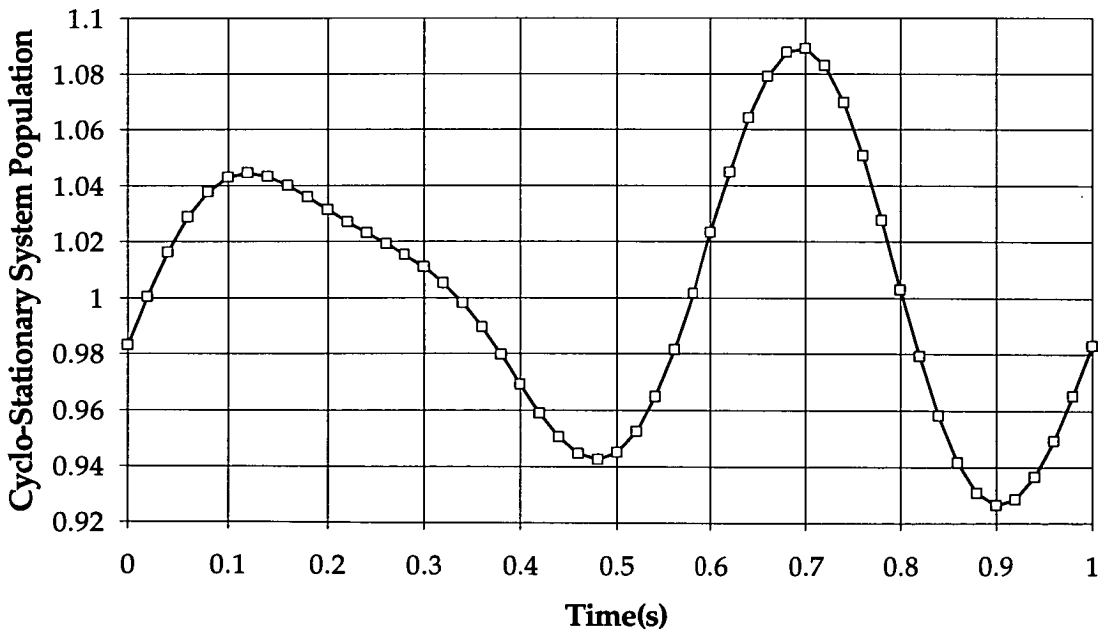


Figure 8.4: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = \pi/4$, $\phi_2 = -\pi/4$

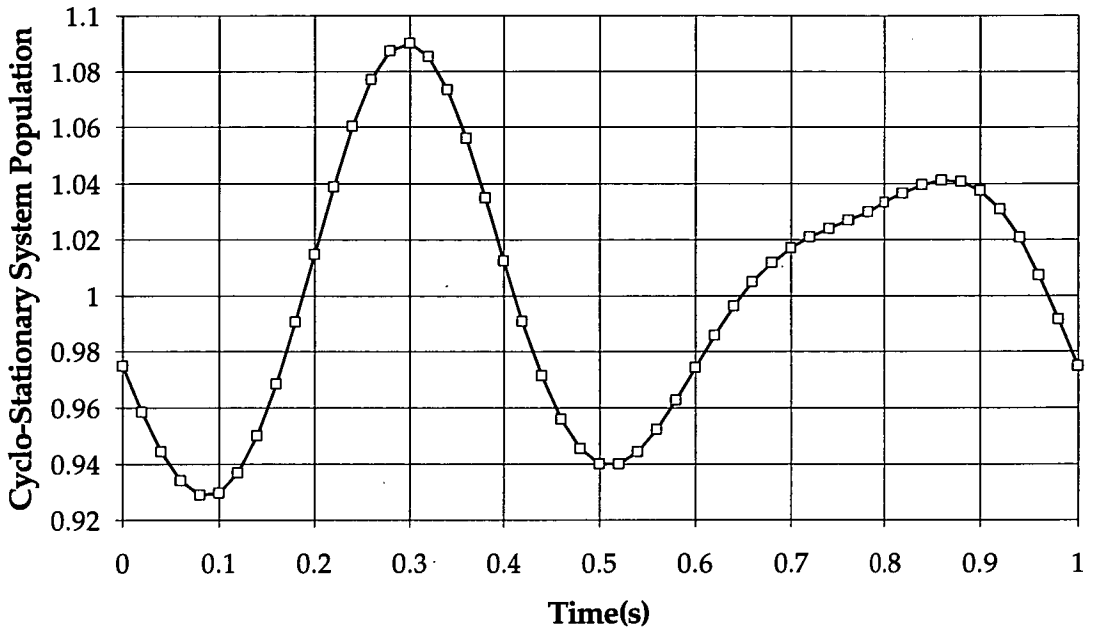


Figure 8.5: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = -\pi/4$, $\phi_2 = \pi/4$

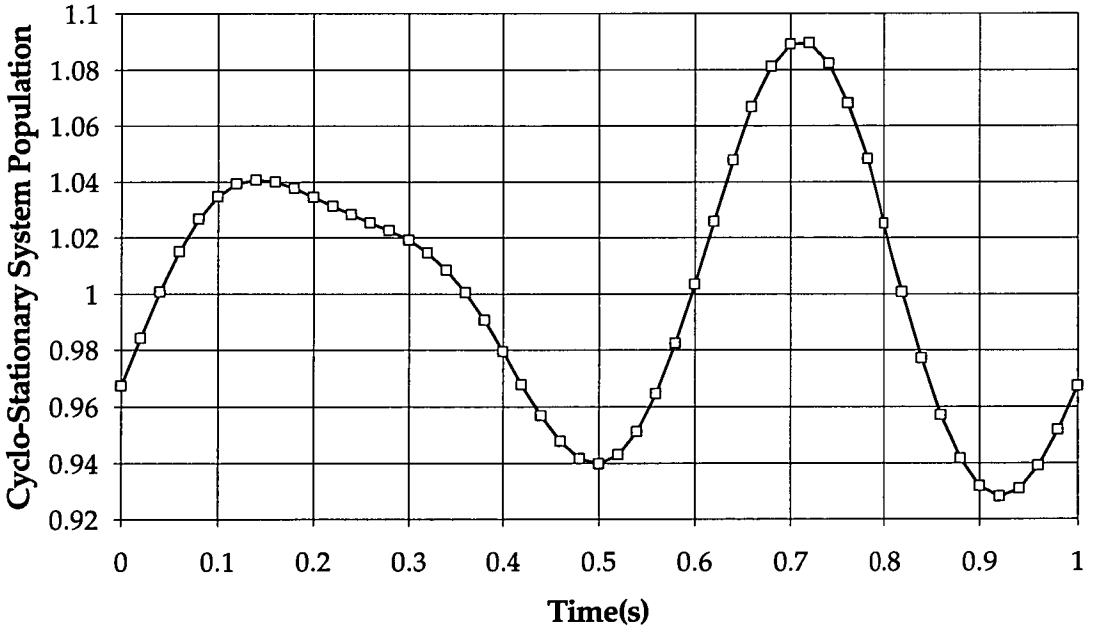


Figure 8.6: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = \pi/6$, $\phi_2 = -\pi/3$

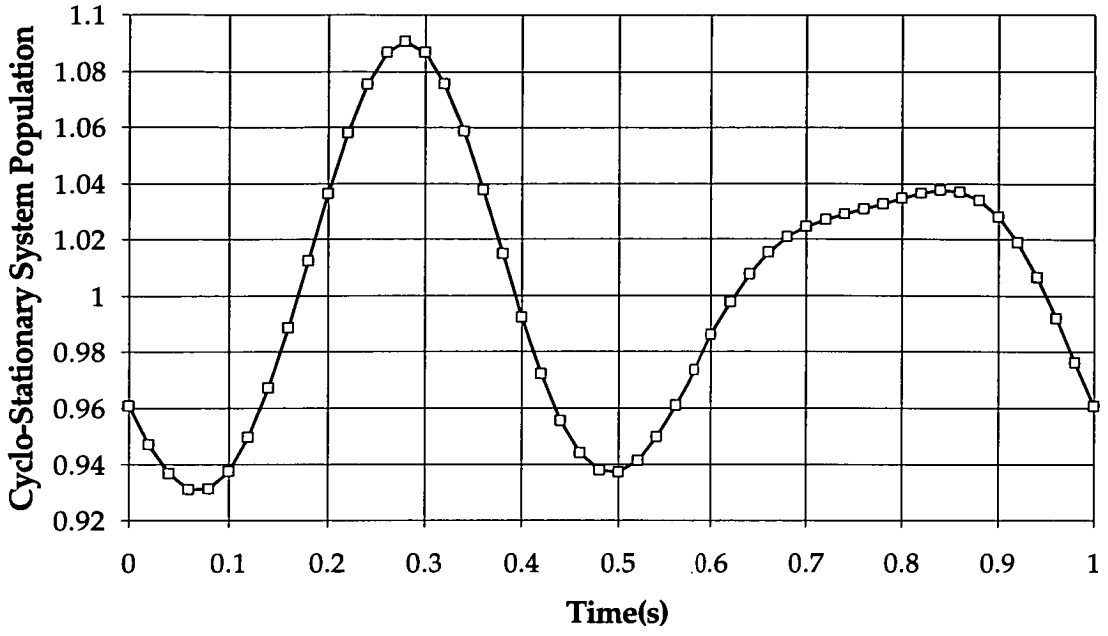


Figure 8.7: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = -\pi/6$, $\phi_2 = \pi/3$

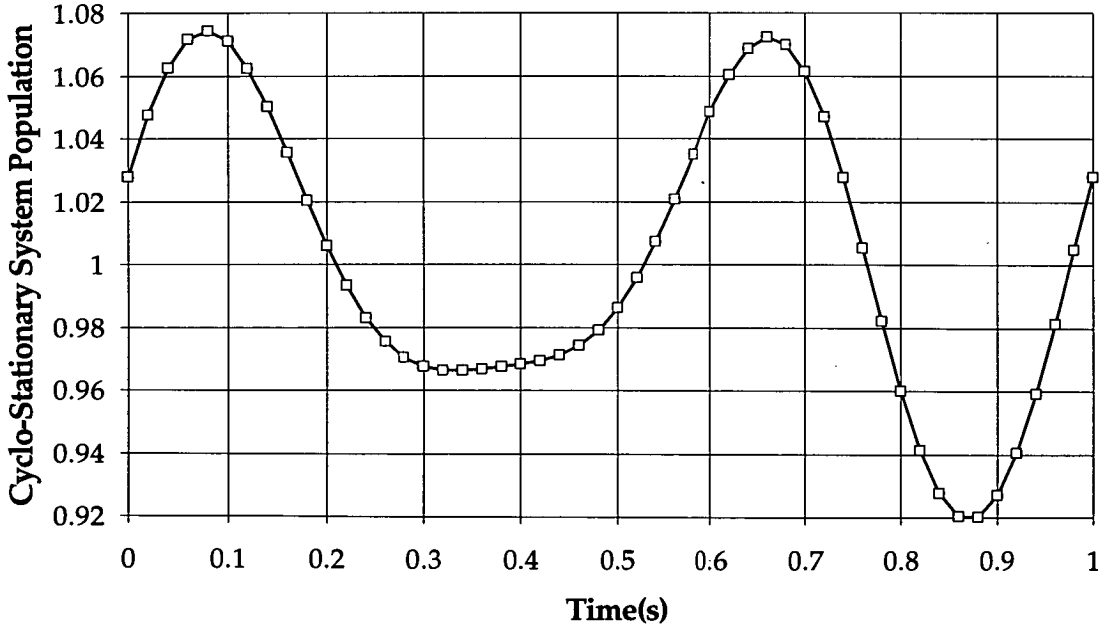


Figure 8.8: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = \pi/2$, $\phi_2 = -\pi/4$

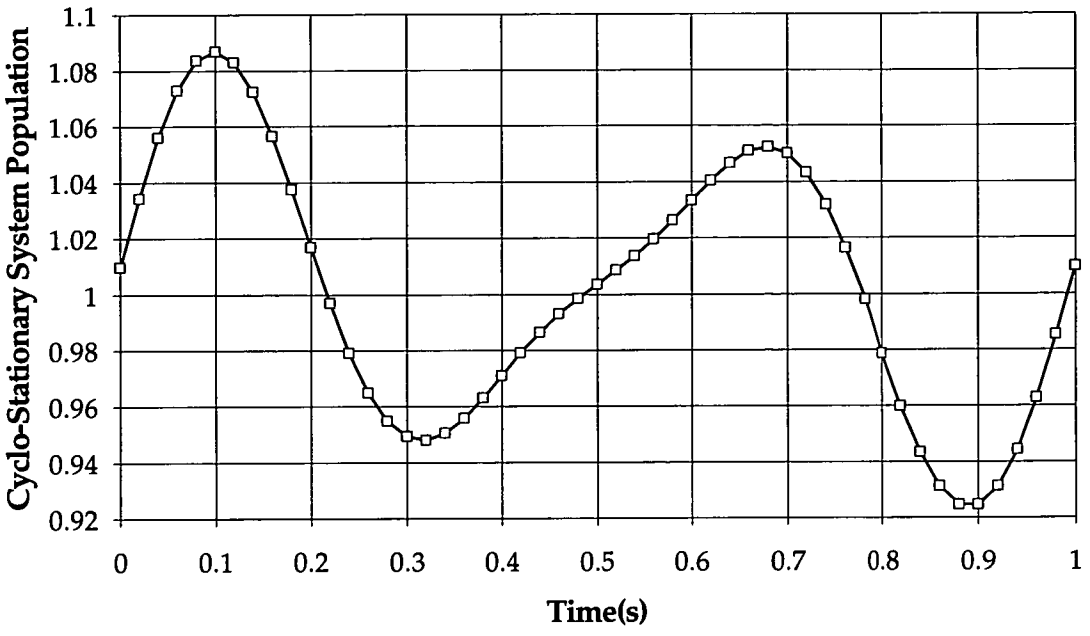


Figure 8.9: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = \pi/2$, $\phi_2 = -\pi/2$

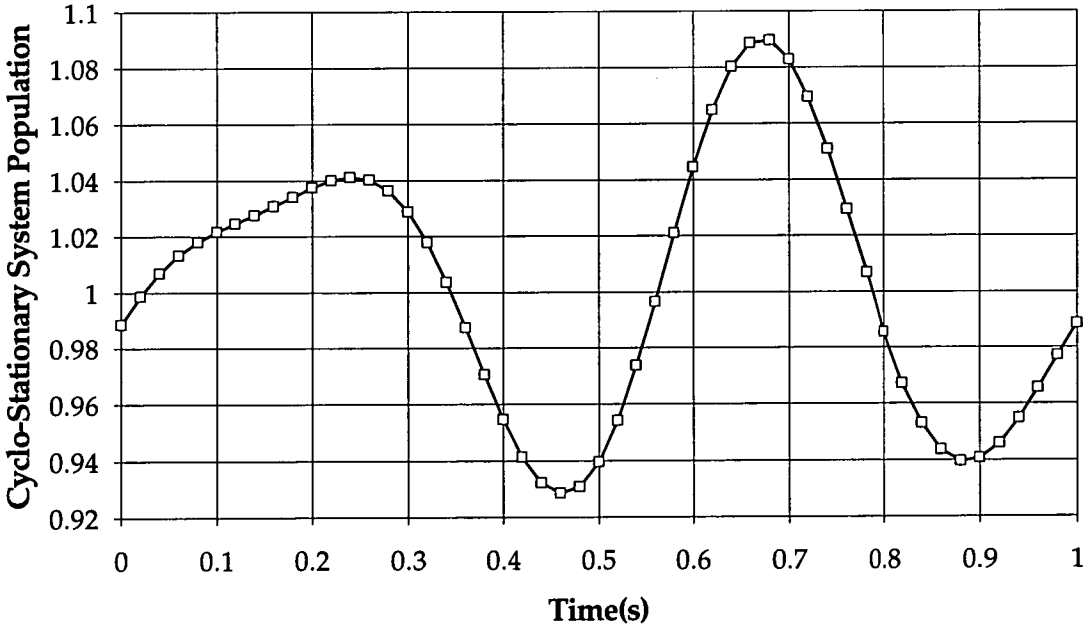


Figure 8.10: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.75$, $\tau = 2.0$, $\gamma = 1.0$, $a_1 = 2$, $a_2 = 3$, $\omega = 2\pi$, $\phi_1 = \pi/4$, $\phi_2 = 0.0$

The numerical solution presented here may also be used to treat the cases where only the service rate has a cyclic nature. For these situations, the time varying term for the arrival rate is zero, i.e. $\beta = 0$. An example of this is shown in Figure 8.11.

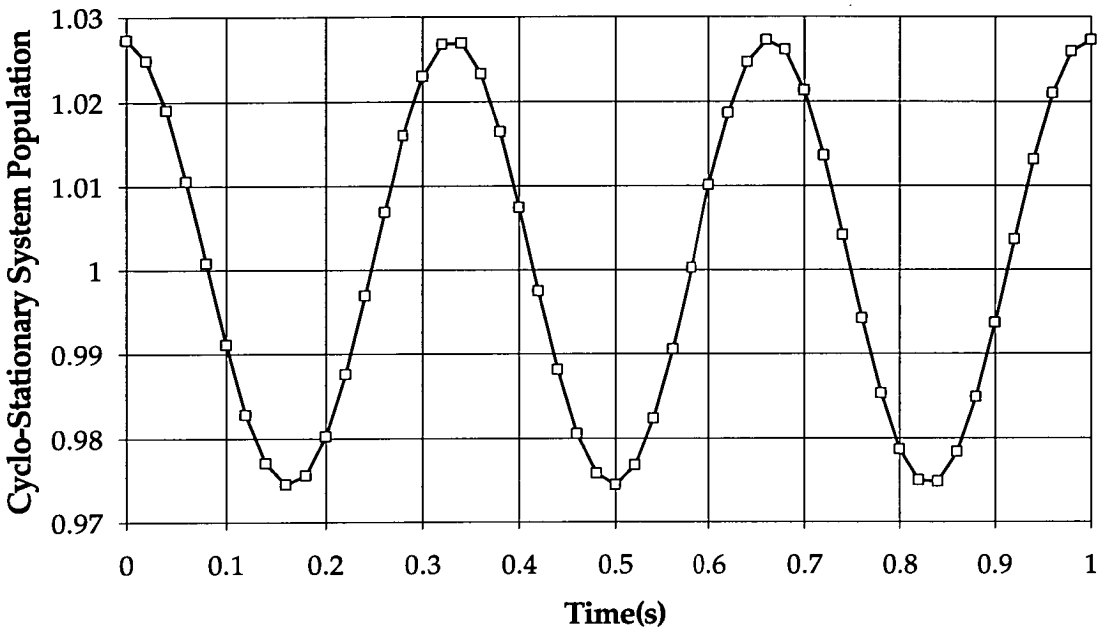


Figure 8.11: Cyclo-stationary system population with $\alpha = 1.0$, $\beta = 0.0$, $\tau = 2.0$, $\gamma = 1.0$, $a_2 = 3$, $\omega = 2\pi$, $\phi_2 = 0.0$

8.3 Generalised Periodic Input & Periodic Output

In this section we consider the situation where both the arrivals and the service rate are cyclo-stationary, but, the mean arrival rate and the mean service rate have arbitrary shapes in their cycles instead of being restricted to sinusoidal shapes. Due to the cyclic nature of the mean arrival rate and the mean service rate, each of them may be expressed as a Fourier series, i.e.

$$\lambda(t) = \sum_{i=-\infty}^{\infty} \beta_i e^{ji\omega_1 t} \tag{8.18}$$

$$\mu(t) = \sum_{i=-\infty}^{\infty} \gamma_i e^{ji\omega_2 t} \tag{8.19}$$

where β_i and γ_i are given by

$$\beta_i = \frac{1}{T}\int_{-T/2}^{T/2} \lambda(t)e^{-ji\omega_1 t}dt \tag{8.20}$$

$$\gamma_i = \frac{1}{T}\int_{-T/2}^{T/2} \mu(t)e^{-ji\omega_2 t}dt \ . \tag{8.21}$$

Now, the results of our studies on the sinusoidal periodic arrival rate and the sinusoidal periodic service rate may be extended for this case. It is implied that a numerical solution for this type of queueing systems is attainable if the fundamental frequencies of the input rate and the output rate of the queue are multiples of a frequency, say ω. Arbitrary phase shifts are also allowed. Therefore, the arrival rate and the service rate for this system may be rewritten as

$$\lambda(t) = \sum_{i=-\infty}^{\infty} \beta_i e^{ji(a_1\omega t+\phi_1)} \tag{8.22}$$

$$\mu(t) = \sum_{i=-\infty}^{\infty} \gamma_i e^{ji(a_2\omega t+\phi_2)} \tag{8.23}$$

where a_1 and a_2 are integers and ϕ_1 and ϕ_2 are phase shifts in radians. Substituting equations (8.22), (8.23) and (8.7) into equation (8.6) gives

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} \ = \ & -\Big(\sum_{i=-\infty}^{\infty} \beta_i e^{ji(a_1\omega t+\phi_1)} + \sum_{i=-\infty}^{\infty} \gamma_i e^{ji(a_2\omega t+\phi_2)}\Big) \sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t} \\ & + \sum_{i=-\infty}^{\infty} \beta_i e^{ji(a_1\omega t+\phi_1)} \sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t} \\ & + \sum_{i=-\infty}^{\infty} \gamma_i e^{ji(a_2\omega t+\phi_2)} \sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t}\end{aligned} \tag{8.24}$$

or

$$\begin{aligned}\sum_{k=-\infty}^{\infty} jk\omega c_{k,n}e^{jk\omega t} \ = \ & -\sum_{i=-\infty}^{\infty} \beta_i e^{ji\phi_1}e^{jia_1\omega t} \sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t} \\ & - \sum_{i=-\infty}^{\infty} \gamma_i e^{ji\phi_2}e^{jia_2\omega t} \sum_{k=-\infty}^{\infty} c_{k,n}e^{jk\omega t} \\ & + \sum_{i=-\infty}^{\infty} \beta_i e^{ji\phi_1}e^{jia_1\omega t} \sum_{k=-\infty}^{\infty} c_{k,n-1}e^{jk\omega t} \\ & + \sum_{i=-\infty}^{\infty} \gamma_i e^{ji\phi_2}e^{jia_2\omega t} \sum_{k=-\infty}^{\infty} c_{k,n+1}e^{jk\omega t} \ .\end{aligned} \tag{8.25}$$

Equation (8.25) implies that for $n \neq 0$:

$$\begin{aligned} jk\omega c_{k,n} &= -\sum_{i=-\infty}^{\infty} \beta_i e^{ji\phi_1} c_{k-ia_1,n} - \sum_{i=-\infty}^{\infty} \gamma_i e^{ji\phi_2} c_{k-ia_2,n} \\ &\quad + \sum_{i=-\infty}^{\infty} \beta_i e^{ji\phi_1} c_{k-ia_1,n-1} + \sum_{i=-\infty}^{\infty} \gamma_i e^{ji\phi_2} c_{k-ia_2,n+1} \end{aligned} \tag{8.26}$$

and for $n = 0$:

$$jk\omega c_{k,n} = -\sum_{i=-\infty}^{\infty} \beta_i e^{ji\phi_1} c_{k-ia_1,n} + \sum_{i=-\infty}^{\infty} \gamma_i e^{ji\phi_2} c_{k-ia_2,n+1} \; . \tag{8.27}$$

This is a recurrence relationship for coefficient $c_{k,n}$ in terms of an infinite number of entries in the array of Fourier series coefficients. In order to solve this problem numerically, the arrival rate and the service rate have to be approximated with a finite number of harmonics. Let l_1 and l_2 denote the limits on the number of harmonics used for approximating the arrival rate and the service rate respectively. Equation (8.26) can therefore be approximated to

$$\begin{aligned} jk\omega c_{k,n} &= -\sum_{i=-l_1}^{l_1} \beta_i e^{ji\phi_1} c_{k-ia_1,n} - \sum_{i=-l_2}^{l_2} \gamma_i e^{ji\phi_2} c_{k-ia_2,n} \\ &\quad + \sum_{i=-l_1}^{l_1} \beta_i e^{ji\phi_1} c_{k-ia_1,n-1} + \sum_{i=-l_2}^{l_2} \gamma_i e^{ji\phi_2} c_{k-ia_2,n+1} \end{aligned} \tag{8.28}$$

and for $n = 0$ it will reduce to

$$jk\omega c_{k,n} = -\sum_{i=-l_1}^{l_1} \beta_i e^{ji\phi_1} c_{k-ia_1,n} + \sum_{i=-l_2}^{l_2} \gamma_i e^{ji\phi_2} c_{k-ia_2,n+1} \; . \tag{8.29}$$

8.4 Summary

In this chapter the method of analysis presented in Chapter 7 has been extended successfully to cater for queueing analysis of systems that have periodic variations in both the arrival rate and the service rate of their traffic. An example of periodic service rate is cyclic suspension of service while other traffic is handled.

Initially a simple case was considered where both the arrival rate and the service rate had sinusoidal variations. The conditions for which numerical results are attainable were investigated. The analysis indicated that when only one of

the input rate or the output rate is periodic then a numerical solution may be found regardless of the shape or the frequency of the periodic variation. For the case where both the input rate and the output rate of the queue are periodic, a numerical solution may always be found except in the unusual situation where the frequency components of the input rate and the output rate have no common divisor. It has been observed that if both the arrival and the service rates are cyclo-stationary, then their phase shifts can affect the performance results considerably. Finally, the analysis was extended for generalised shapes of cyclo-stationary arrival and service rates.

Chapter 9

Summary and Future Extensions

9.1 Introduction

This thesis has encompassed a variety of research areas related to performance analysis of Broadband Integrated Services Digital Networks. In this final chapter a summary of the highlights of this work and the important results obtained from them are described. Also, a number of extensions to the work presented in this thesis are outlined for future research.

9.2 An Overview

This thesis has given an introduction to B-ISDNs. This introduction has included a description of the earlier networks and technologies that have evolved into the introduction of B-ISDNs. It has also described the service classes for B-ISDN, the B-ISDN protocol reference model, the Asynchronous Transfer Mode (ATM), and a detailed description of traffic control in B-ISDN.

Next, the thesis has considered the problem of dynamic allocation of capacity at an access node to a mixture of two types of service, one type being delay sensitive and the other type loss sensitive. Several strategies have been considered and different methods of analysis have been employed to obtain performance measures for the access node, under these strategies.

Simulation tools have been used to study an access node where three types of service (interactive data, VBR video, and interactive images) are multiplexed in an ATM environment. A simple strategy has been implemented for multiplexing the traffic generated from these services. The effects of the ratio of the link bit rate to source peak bit rates have been studied. A particular method of packaging the video information into cells with high and low priorities has been employed to study the effects of introducing priority to the cell stream of the video traffic on the performance of the access node.

A particular method for dynamic allocation of capacity to a mix of fixed bit rate and queueable variable bit rate services has been examined. Some performance results for both classes of traffic have been generated from different methods of analysis. These results have been confirmed by simulation.

The thesis has then considered performance models for the traffic generated from VBR video services. A literature review has been presented that summarizes the major models proposed for modelling VBR video traffic. Three video traffic models have been developed based on the concept of an underlaying hidden Markov model. The performance of these models have been studied for different link utilisations, and comparisons made with performance results of the actual video traffic. Some correlation studies have also been undertaken for the traffic generated from the actual VBR video traffic and for the traffic generated by the models.

As its last topic, the thesis has considered performance modelling for queueing systems that have cyclo-stationary variations in the arrival rate and/or service rate of their traffic. These queues have direct application in the analysis of some communication & computer networks. A method of analysis has been developed that uses Fourier series to arrive at a numerical solution for calculating the probabilities of various states of the queueing system as a function of time. This method is easy to implement and can calculate, as a function of time, all the usual performance parameters for a whole cycle of the system under consideration. Several examples have been solved which prove this method to be a powerful tool for the analysis of cyclo-stationary queueing systems. Effects of the

truncation of the Fourier series coefficients have been considered. Effects of the frequency of variation of the traffic on the performance of the system have been studied. For the case where both the arrival rate and the service rate are cyclic, the conditions for which a numerical solution is attainable have been investigated.

9.3 Summary of Results

In this section the important results and highlights of this thesis will be outlined. The results will be given under three headings, corresponding to the three streams of performance modelling that have been considered in this thesis.

9.3.1 Performance Modelling of Access Control

Dynamic Allocation of Capacity in TDM & ATM Environments

Initially several control strategies were studied for an access node multiplexer that serves wideband (WB) and narrowband (NB) traffic in a synchronous TDM environment. The strategies considered were MBNSD (movable boundary with no sorting of the channel allocations of the digital pipe), MB (movable boundary with sorting of channels allocations of the digital pipe), and MBP (movable boundary with pre-emption).

These strategies were analysed through different methods including simulation, an approximate Markov chain analysis (iterative), and a decomposition method of matrix-geometric analysis. Results from various methods of analysis were found to be almost identical. NB traffic received the most favourable treatment under the MBNSD strategy, and the least favourable treatment under the MBP strategy. The order of strategies for favouring WB traffic was MBP first, MB second and MBNSD third. A combined performance measure indicated that MBNSD was a better strategy overall. Interestingly enough, MBNSD is also the easiest strategy to implement because it does not require the access node to have the capability of reassigning the channel allocations for NB calls in progress.

The MBP strategy was then modified for an ATM environment and the performance of the system was analysed using simulation, an approximate Markov chain analysis, and an iterative method of matrix geometric analysis. For the same set of traffic parameters, statistical multiplexing showed an improvement in the performance of the NB traffic as compared to synchronous TDM multiplexing. For the case studied, the factor of improvement in the delay of the NB traffic was between 2.5 to 6 depending on the NB traffic load.

Among different methods of analysis, the decomposition method of matrix-geometric was found to be the fastest. None of the problems investigated needed the slower, but more accurate iterative techniques.

Mixing Interactive Images, Data & Video Traffic

An access node in an ATM network that serves interactive data, VBR video and interactive images was studied using simulation. It was found that under the condition that all traffic types have burst bit rates much smaller than the output link bit rate, reasonably high utilisations can be achieved under a null access strategy that treats all traffic types and all cells equally.

The ratio of the link bit rate to the source bit rates was then reduced and it resulted in significant degradation in the performance of the access node.

A particular cell packaging scheme was then used to produce two priority levels for the ATM cells generated from the video signal. The access control scheme was modified so that when the buffer occupancy was above a threshold level, all the low priority video cells were discarded. It was found that with low utilisations, introducing priority levels for video cells did not improve the performance of the access node as expected. In fact, it degraded the performance because under the same coding scheme, priority encoded video has a higher bit rate compared to non-priority encoded video. As the utilisation increases however, the advantages of priority encoding the video traffic outweigh its disadvantage of higher bit rate, and it actually improves the cell delays at the access node.

Mixing Queueable VBR Traffic with CBR Traffic

An ATM access node was considered that serves a range of CBR services and queueable VBR traffic. A strategy was proposed for sharing the capacity between these services. A Markov process was used to calculate performance measures for the CBR traffic. Then, assuming that the mean arrival rate of the CBR calls is much smaller than the mean arrival rate of the VBR cells and given the fixed size of cells, the Pollaczek-Khinchin mean value formula for an M/G/1 system was used to calculate the mean queue size for the VBR traffic.

An imbedded Markov chain was then used to analyse the performance measures for the VBR traffic, i.e. to calculate the variance of the queue size for the VBR cells as well as the mean queue size (which had also been calculated using Pollaczek-Khinchin formula). The performance measures obtained for the VBR traffic from the two methods and from simulation were compared and were found to be in good agreement. Simulation was also used to confirm the results obtained for the CBR and the VBR traffic.

9.3.2 Performance Modelling of Video Traffic

The suitability of hidden Markov models for modelling the cell stream generated from a VBR video codec was investigated. These models were based on dividing each video frame into several fixed size blocks, grouping a number of blocks into a subframe and then assigning a mode to each subframe depending on the number of ATM cells generated from that subframe.

For the basic HMM video model, the actual VBR video traffic was used to generate an intermode transition probability matrix for modelling the transitions between modes. Furthermore, intramode cell generation transition probability matrices were generated to model the cell generation within each mode. The model was used to generate traffic and the traffic, in the form of ATM cells, was fed to a single server queue. Some results were generated for the mean and the variance of the queue population. Probability density functions (pdfs) for the number of ATM cells per subframe were also generated . These results were

compared with those from the actual video traffic. It was found that the size of the subframe has a significant effect on the accuracy of the model, particularly at high utilisations.

The original HMM model was simplified to HMD model, where intramode matrices were no longer used to model the cell generation within each mode. Instead, a deterministic number of cells, randomly distributed over the blocks of a subframe, were generated in each mode. It was found that the pdf results of the model had improved and were almost identical to the pdf results of the actual video traffic. However, for high utilisations, some loss of accuracy had been incurred in the results for the mean and the variance of the queue population.

Another model (designated HMDL) was developed which was similar to HMD except that no block within a subframe could generate more ATM cells than the maximum observed value for the real video traffic. This resulted in a significant gain in the accuracy of the model. The HMDL model is simpler than the HMM model in that it requires much fewer number of parameters for its complete specification, but it can still track the original video data very closely for queueing purposes. The results of the HMM, the HMD and the HMDL models indicate that hidden Markov models can successfully be applied in the modelling of variable bit rate video services.

Some correlation studies were undertaken for the video traffic and it was found that strong cyclic variations were present in the cell stream of the video traffic particularly at the frame rate and the line rate.

9.3.3 Performance Modelling of Cyclo-Stationary Queueing Systems

The starting point of this work was a system consisting of a single server queue with random arrivals and random service rate, but, with a mean arrival rate that was cyclic (i.e. the arrivals were cyclo-stationary). The shape of the arrival rate

was taken to vary sinusoidally around a fixed term. An analysis method was presented based on the idea of using a Fourier series with complex coefficients to describe the cyclo-stationary probabilities of being in different states of the system. All other performance measures can be calculated from these coefficients.

This analysis resulted in a non-linear, complex recurrence relationship that described each Fourier series coefficient in terms of its neighbouring coefficients. We found that the convergence of the Fourier series coefficients was dependent on how the recurrence relationship was applied to them. A particular method based on recurring on the entries of the array of Fourier series coefficients on a block by block basis resulted in very fast convergence.

The effect of truncating the Fourier series coefficients on the accuracy of the results was considered. The conclusion was that in order to select an appropriate size for the array of the Fourier series coefficients, the smallest probabilities of interest should be estimated, say P_1 (e.g. in a packet switched network, this may be estimated from such parameters as the acceptable cell loss ratio). Then the M/M/1 queueing results should be used to find the smallest system population, say n_1, that conforms to $P(n_1) \leq P_1$. The dimensions of the array of Fourier series coefficients should be initially several times larger than n_1 and should be determined after a few trials.

The effect of the frequency of the arrival rate on the performance of the queue was considered. With all other parameters fixed, it was observed that the variation in the cyclo-stationary system population decreased as the frequency of variation of the arrival rate increases. This is in line with the argument that with lower frequencies the arrival rate spends a longer duration around each extreme in each cycle, hence giving more time to the system to adjust to the variations of the arrival rate.

The method of analysis was extended to cater for arrivals that are cyclo-stationary, but have a mean arrival rate that has an arbitrary shape in each cycle. The arrival rate itself then had to be described as a truncated Fourier series. The analysis

resulted in a recurrence relationship which contained many more terms. The computation time increased considerably, but the results still converged satisfactorily. As an example, a case in which the arrival rate switched between two mean rates (square waveform) was analysed and performance results were generated.

The next extension to this work was to consider a queueing system where both the arrivals and the service rate have cyclo-stationary variations. This queueing system was also analysed using similar techniques. Initially a simple case was considered where both the arrival rate and the service rate had cyclo-stationary variations, and were sinusoidal in shape. The conditions for which numerical results are attainable were investigated. The analysis indicated that when only one of the input rate or the output rate is periodic then a numerical solution is always attainable regardless of the shape or the frequency of the periodic variation. For the case where both the input rate and the output rate of the queue are periodic, a numerical solution may always be found except in the unusual situation where the frequency components of the input rate and the output rate have no common divisor. For this case it was also observed that phase shifts in the arrival rate and/or service rate can affect the performance results considerably. Finally, the analysis was extended for generalised shapes of cyclo-stationary arrival and service rates.

9.4 Suggestions for Future Extensions

One of the possible extensions to the work presented on traffic modelling of VBR video services is to analytically quantify some queueing results for the video traffic from the parameters of the hidden Markov model.

If the mode definitions of the hidden Markov model are modified such that a large number of cells are generated in each mode, then it may be possible to approximate the generation of ATM cells on a mode by mode basis by an M/D/1/N queue. If the arrival of ATM cells in each mode could be approximated by a poisson process with a particular mean value, and if the number of cells generated in each mode is large enough so that the system may be assumed to reach

steady state during each mode, then the statistics of the system may be obtained by solving the M/D/1/N system for different modes and then conditioning over the range of the modes. The steady state probabilities of different modes may be obtained from the intermode transition probabilities matrix.

A further research on hidden Markov models would be to find its limitations in relation to the bit rate of the VBR video traffic, i.e. to find if these models can successfully be applied to very low bit rate video traffic. We expect that for very low bit rate video traffic, the accuracy of the hidden Markov models will be reduced because the subframes may generate less than one ATM cell on average and there will be a lot of overlap between consecutive subframes. However, because video compression and coding is not within the scope of this thesis we have not tried the hidden Markov models for different video bit rates.

Another extension to the work presented in this thesis could be to try other approaches for finding solutions to some of the cyclo-stationary queueing problems that were considered in Chapter 7 and Chapter 8. One approach to this problem could be to use fluid flow models. These models treat the queueing system as a liquid reservoir. If $q(t)$ is the population of the system at time t and if $\lambda(t)$ and $\mu(t)$ are the arrival rate into the queue and the departure rate out of the queue respectively, then the rate of change in the population of the queue, $\dot{q}(t)$, may be written as

$$\dot{q}(t) = \lambda(t) - \mu(t) \ . \tag{9.1}$$

This equation is referred to as the fluid flow equation and has many applications in the analysis of queueing systems.

Another extension for future research is in the area of design and analysis of congestion control schemes for B-ISDN. From the published research in the area of access control and congestion control in B-ISDN, two approaches may be identified. Some researchers have considered the problem in a broad sense [142, 119, 143, 144], only describing some general guidelines for various aspects of congestion control, without proposing and analysing detailed schemes. There are other works which more explicitly describe and analyse some aspects of traffic control

[145, 146, 147], but most of these only take into account an overly simplified set of traffic parameters in their design. Some researchers [148, 28, 149] address such issues as the effects of correlation present in the pattern of cell arrivals from particular services, but these considerations are not generally taken into account in the design of traffic control schemes. Furthermore, many of the proposed traffic control strategies [150, 151] are too service specific and although they may work well with a particular traffic type, they may not cope as well with other service types with different grade of service requirements from the network. Designing such service specific traffic control schemes may not be appropriate in the real B-ISDN where there is a wide range of services which are very different in their traffic characteristics and QOS requirements. It is highly desirable to have traffic control schemes that are more intelligent than most schemes so far proposed, many of which only consider mean and peak bit rates, and perhaps some measure of burstiness.

Such intelligence should be built into the frame work of the traffic control schemes without imposing too much processing burden on the network. It is important from the network operator's point of view to be able to optimize the network based not only on the current loading and the current number of service requests, but also based on contingent future demand for the network resources on a time scale of the order of, say, one call duration.

Therefore, an area for further research may be to investigate the performance sensitivity of the network to various traffic parameters and to develop (based on sensitivity studies) more intelligent access control schemes that can be applied to a wider range of services. Ideally, it should be a unified access control scheme based on the idea of a cost function which may be subject to optimisation for fine tuning. This approach means that access control scheme can also be trained by the dynamics of the network.

Appendix A

Markov Chains & Markov Processes

Two of the most important concepts in queueing theory are Markov chains and Markov processes. In performance analysis of high speed telecommunication networks, Markov chains are commonly used to model individual multimedia sources, to capture their essential time-autocorrelation properties [152]. This appendix briefly covers Markov chains and Markov processes [93, 109, 153, 154] since they have been used frequently throughout this thesis.

A.1 Elementary Theory of Markov Chains

A *Stochastic process* is a function of time whose values are random variables, for example the number of people sitting in a movie theatre as a function of time. A *Markov process* is a stochastic process that the probability distribution of its future development depends only on the present state and not on how the process arrived in that state [93].

If the state space, I, is taken to be discrete, then the Markov process is known as *Markov chain.* Furthermore, if the parameter space, T, is also discrete, then we have a *discrete-parameter Markov chain.* The transition probability matrix of a Markov chain is defined as $\mathbf{P} = [p_{ij}]$ where p_{ij} is defined as [109]:

$$p_{ij} = P[X_n = j | X_{n-1} = i] \ .$$

This definition assumes that the Markov chain is homogeneous. Let us define $\pi_j^{(n)}$ as the probability of finding the system in state E_j at the n^{th} step, i.e.:

$$\pi_j^{(n)} = P[X_n = j] \ .$$

Furthermore, if the Markov chain is irreducible and aperiodic homogeneous, then the limiting probabilities

$$\pi_j = \lim_{n\to\infty} \pi_j^{(n)}$$

always exist and are independent of the initial state probability distribution. Moreover:

- *either* all states are transient or all states are recurrent null in which case $\pi_j = 0$ for all j and there exists no stationary distribution,

- *or* all states are recurrent non-null and then $\pi_j > 0$ for all j in which case the set $\{\pi_j\}$ is a stationary probability distribution and the quantities π_j are uniquely determined through the following equations:

$$1 = \sum_i \pi_i \tag{A.1}$$

$$\pi_j = \sum_i \pi_i p_{ij} \ . \tag{A.2}$$

If we further define the invariant probability vector π as:

$$\pi = [\pi_0, \pi_1, \pi_2, ...]$$

then we may rewrite the set of equations in (A.2) as:

$$\pi = \pi \mathbf{P} \ . \tag{A.3}$$

It is important to note that in (A.3), always one of the equations will be dependent on the others and it is therefore necessary to introduce the constraint given in equation (A.1) in order to solve the system of linear equations.

A.2 Treatment of Higher Order Markov Processes

One of the major problems in dealing with multidimensional Markov chains is that the size of the transition probability matrix grows exponentially with the dimension of the Markov chain. As an example let us take the 1D case and assume that there exist m possible states. Therefore, the transition probability matrix is a $m \times m$ matrix. Now if we had a 2D Markov chain with m_1 and m_2 states in each dimension, the transition probability matrix would be $m_1 m_2 \times m_1 m_2$. In theory, so long as the state-space is finite, one should be able to treat a multidimensional Markov chain like a 1D case and use equations (A.1) and (A.2) to find the invariant probability vector π.

In reality however, computational difficulties are encountered for solving large set of simultaneous equations. Furthermore, in many practical situations, the state-space becomes infinite which means that equations (A.1) and (A.2) can not be solved directly and other methods have to be used. Transform techniques, such as the moment-generating functions approach[109], are often used in the study of queueing systems. These methods require very complex analysis and are of substantial mathematical difficulty. For example, in the case of moment generating functions the difficulty is in the evaluation of a large number of boundary terms which themselves may require the determination of multiple zeroes of a high degree polynomial or a transcendental function.

One of the methods which offers greater flexibility in analysis of higher order Markov processes is the *Matrix-Geometric Solutions* method. This method has been introduced in the last decade by Neuts [155] and is based on an iterative algorithm for finding the invariant probabilities.

A.3 Matrix-Geometric Solutions Method

In this section we have extracted those definitions and theorems given in [89] that relate to the work presented in this thesis. Two of the most relevant instances of

Markov chains are the embedded Markov chains of the elementary M/G/1 and G/M/1 queues (p. 1 of [89]). The block-partitioned matrices for these chains are respectively given as:

$$\bar{P}_1 = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & B_4 & \dots \\ C_0 & A_1 & A_2 & A_3 & A_4 & \dots \\ 0 & A_0 & A_1 & A_2 & A_3 & \dots \\ 0 & 0 & A_0 & A_1 & A_2 & \dots \\ 0 & 0 & 0 & A_0 & A_1 & \dots \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{bmatrix} \tag{A.4}$$

$$\bar{P}_2 = \begin{bmatrix} B_0 & A_0 & 0 & 0 & 0 & \dots \\ B_1 & A_1 & A_0 & 0 & 0 & \dots \\ B_2 & A_2 & A_1 & A_0 & 0 & \dots \\ B_3 & A_3 & A_2 & A_1 & A_0 & \dots \\ B_4 & A_4 & A_3 & A_2 & A_1 & \dots \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{bmatrix} \tag{A.5}$$

where the elements $A_v, v \geq 0, B_v, v \geq 1, B_0$ and C_0 are finite, non-negative matrices of dimensions $m * m, n * m, n * n$ and $m * n$ respectively[89]. Since matrix $\bar{P}_2$ is stochastic, we clearly have

$$B_k e + \sum_{v=0}^{k} A_v e = e \quad ; \text{for} \quad k \geq 0 \tag{A.6}$$

where e is a column vector with all its components equal to one. Although no easily verifiable criterion for irreducibility is available, we assume that the Markov chain $\bar{P}_2$, which from now on will simply be denoted by $\bar{P}$, is irreducible. This is nearly always the case in well-formulated practical models. The structure of the matrix $\bar{P}$ implies that in a single transition the chain can move upwards only to the next higher level. In moving from the state (i, j) to a state $(i + k, v)$, with $k \geq 1$, the chain must visit all intermediate levels at least once.

We shall now proceed to study the conditions under which the Markov chain $\bar{P}$ is positive recurrent. The invariant probability vector p of $\bar{P}$, in the positive

recurrence case, is the unique solution to the infinite system of equations

$$p\bar{P} = p \tag{A.7}$$

$$pE = 1 \tag{A.8}$$

where E is a column vector whose components are all e's. The invariant probability vector p is partitioned as

$$p = [p_0, p_1, p_2, p_3, ...]$$

where the row vectors p_k are given by

$$p_k = [p_{k,0}, p_{k,1}, p_{k,2}, ...]$$

Equations (A.7) & (A.8) may also be rewritten as (p. 7 of [89]):

$$\begin{aligned} p_0 &= \sum_{v=0}^{\infty} p_v B_v \\ p_k &= \sum_{v=0}^{\infty} p_{k+v-1} A_v \quad , \quad \text{for } k \geq 1 \\ \sum_{k=0}^{\infty} p_k e &= 1 \ . \end{aligned} \tag{A.9}$$

We now have to define the *taboo probability* ${}_{\mathbf{i}}P^{(n)}_{i,j;i+k,v}$ as the probability that starting in the state (i,j), the chain reaches (i+k,v) at time **n** without returning to any states in level **i** at any time in between. This probability is defined for $n \geq 0, i \geq 0, k \geq 1, 1 \leq j, v \leq m$ and is equal to zero for $n < k$. More important is the fact that the particular structure of $\bar{P}$ makes the value of this taboo probability independent of i for all $n \geq 0, k \geq 1, 1 \leq j, v \leq m$. The taboo probability ${}_{\mathbf{i}}P^{(n)}_{i,j;i+k,v}$ depends solely on the structure of the submatrix of $\bar{P}$ obtained by deleting all rows and columns with indices (r, j'), $r \leq 0$, $1 \leq j' \leq m$. These submatrices are identical for all $i \geq 0$.

We define $R^{(k)}_{jv}$ as the expected number of visits to the state $(i+K, v)$ before the first return to the level **i**, given that the chain $\bar{P}$ starts in the state (i, j). In other words for $k \geq 1$, $i \geq 0$, $1 \leq j \leq m$, $1 \leq v \leq m$ we have

$$R^{(k)}_{jv} = \sum_{n=0}^{\infty} {}_{\mathbf{i}}P^{(n)}_{i,j;i+k,v} \tag{A.10}$$

The square matrix with elements $R_{jv}^{(k)}$, $1 \leq j, v \leq m$, is denoted by $R^{(k)}$ for $k \geq 1$. By agreement $R^{(0)}$ is set to I, the identity matrix. Also $R^{(1)}$ is simply denoted as R and is called the *rate matrix* of chain $\bar{P}$. The following lemma is stated without proof:

Lemma 1 *If the Markov chain $\bar{P}$ is positive recurrent then the matrices $R^{(k)}$, $k \geq 1$ are finite.*

Lemma 2 *The matrix $R^{(k)}$ is the k-th power of the matrix R.*

Proof : For $n \geq k+1$ we have

$$\begin{aligned} {}_{\mathbf{i}}P_{i,j;i+k+1,v}^{(n)} &= \sum_{h=1}^{m}\sum_{r=0}^{n} {}_{\mathbf{i}}P_{i,j;i+k,h}^{(r)}\ {}_{\mathbf{i+k}}P_{i+k,h;i+k+1,v}^{(n-r)} \\ &= \sum_{h=1}^{m}\sum_{r=0}^{n} {}_{\mathbf{i}}P_{i,j;i+k,h}^{(r)}\ {}_{\mathbf{i}}P_{i,h;i+1,v}^{(n-r)} \ . \end{aligned} \tag{A.11}$$

The first equality is obtained from the law of total probability, by conditioning on the time r of the last visit to the level $\mathbf{i+k}$ and on the state $(i+k,h)$ of that visit, before the chain reaches the state $(i+k+1,v)$ at time n. The second equality follows from the fact that ${}_{\mathbf{i}}P_{i,h;i+1,v}$ does not depend on $\mathbf{i}$. For $n \leq k$, both sides of A.11 are zero, so that the equality holds for $n \geq 0$. Summation on n now yields

$$\begin{aligned} R_{jv}^{(k+1)} &= \sum_{h=1}^{m}\sum_{n=0}^{\infty}\sum_{r=0}^{n} {}_{\mathbf{i}}P_{i,j;i+k,h}^{(r)}\ {}_{\mathbf{i}}P_{i,h;i+1,v}^{(n-r)} \\ &= \sum_{h=1}^{m}\sum_{r=0}^{\infty} {}_{\mathbf{i}}P_{i,j;i+k,h}^{(r)} \sum_{n'=0}^{\infty} {}_{\mathbf{i}}P_{i,h;i+1,v}^{(n')} \\ &= \sum_{h=1}^{m} R_{jh}^{(k)} R_{hv} \end{aligned} \tag{A.12}$$

so that $R^{(k+1)} = R^{(k)}R$, and hence $R^{(k)} = R^k$, for $k \geq 1$.

Lemma 3 *If the Markov chain $\bar{P}$ is positive recurrent, the matrix R satisfies the equation:*

$$R = \sum_{k=0}^{\infty} R^k A_k \tag{A.13}$$

and is the minimal non-negative solution to the matrix equation

$$X = \sum_{k=0}^{\infty} X^k A_k \ . \tag{A.14}$$

Proof : Clearly ${}_{\mathbf{i}}P^{(1)}_{i,j;i+1,v} = (A_0)_{jv}$. For $n \geq 2$, by conditioning on the state $(i+k,h)$ from which the state $(i+1,v)$ is entered at time n,

$$ {}_{\mathbf{i}}P^{(n)}_{i,j;i+1,v} = \sum_{h=1}^{m}\sum_{k=1}^{\infty} {}_{\mathbf{i}}P^{(n-1)}_{i,j;i+k,h}\,(A_k)_{hv} \ . \tag{A.15} $$

Summation on n from 2 to ∞ then yields

$$ \begin{aligned} R_{jv} - (A_0)_{jv} &= \sum_{h=1}^{m}\sum_{k=1}^{\infty}\sum_{n'=1}^{\infty} {}_{\mathbf{i}}P^{(n')}_{i,j;i+k,h}\,(A_k)_{hv} \\ &= \sum_{h=1}^{m}\sum_{k=1}^{\infty} R^{(k)}_{jh}(A_k)_{hv} \ . \end{aligned} \tag{A.16} $$

Since $\mathbf{R^{(0)}} = \mathbf{I}$ implies $R^{(0)}_{jj} = 1$,

$$ \begin{aligned} R_{jv} &= \sum_{h=1}^{m}\sum_{k=1}^{\infty} R^{(K)}_{jh}(A_k)_{hv} + R^{(0)}_{jj}(A_0)_{jv} \\ &= \sum_{k=0}^{\infty}\sum_{h=1}^{m} R^{(k)}_{jh}(A_k)_{hv} \end{aligned} \tag{A.17} $$

and therefore the first part of the lemma is proved.

Now consider the sequence {X(N), N $\geq$ 0} of matrices, obtained by performing successive substitutions in $X = \sum_{k=0}^{\infty} X^k A_k$, starting with $X(0) = 0$. It can be shown by induction that the sequence $\{X(N)\}$ is entry-wise nondecreasing and $X(N) \leq R$, for $N \geq 0$. Furthermore, this sequence converges monotonically to a non-negative matrix X^*, which by the dominated convergence theorem, satisfies equation A.13. X^* is called the minimal non-negative solution to equation A.13. It is readily verified that any other non-negative solution X^o, must satisfy $X^* \leq X^o$. In particular, $X^* \leq \mathbf{R}$. We now need to show that $R \leq X^*$. Let us define the matrices $R(N,k)$ with elements

$$ R_{jv}(N,k) = \sum_{n=1}^{N} {}_{\mathbf{i}}P^{(n)}_{i,j;i+k,v} $$

for $k \geq 1, 1 \leq j,v \leq m$. Now, by adding the equations (A.15) for n ranging from 1 to N we obtain

$$ R(N,1) = A_0 + \sum_{k=1}^{\infty} R(N-1,k)A_k \ . \tag{A.18} $$

Also by summing equation (A.11) for n ranging from 1 to $N-1$ we obtain

$$\begin{aligned}
\sum_{n=1}^{N-1} {}_{\mathbf{i}}P^{(n)}_{i,j;i+k+1,v} &= [R(N-1,K+1)]_{jv} \\
&= \sum_{h=1}^{m}\sum_{n=0}^{N-1}\sum_{r=0}^{n} {}_{\mathbf{i}}P^{(r)}_{i,j;i+k,h}\ {}_{\mathbf{i}}P^{(n-r)}_{i,h;i+1,v} \\
&= \sum_{h=1}^{m}\sum_{r=0}^{N-1}\sum_{n=0}^{N-1-r} {}_{\mathbf{i}}P^{(r)}_{i,j;i+k,h}\ {}_{\mathbf{i}}P^{(n)}_{i,h;i+1,v} \\
&\leq \sum_{h=1}^{m}\sum_{r=0}^{N-1} {}_{\mathbf{i}}P^{(r)}_{i,j;i+k,h} \sum_{n=0}^{N-1} {}_{\mathbf{i}}P^{(n)}_{i,h;i+1,v} \\
&= \sum_{h=1}^{m}[R(N-1,k)]_{jh}[R(N-1,1)]_{hv} \qquad \text{(A.19)}
\end{aligned}$$

or in matrix form:

$$R(N-1,k+1) \leq R(N-1,k)R(N-1,1)\ . \qquad \text{(A.20)}$$

By induction, $R(N-1,k) \leq [R(N-1,1)]^k$, for $k \geq 1$. Substituting this in equation (A.18) yields:

$$R(N,1) \leq [R(N-1,1)]^k A_k \qquad \text{for } N \geq 2\ . \qquad \text{(A.21)}$$

We now have that $R(1,1) = A_0 = X(1)$, so that $R(2,1) \leq X(2)$. By induction, equation (A.21) yields that $R(N,1) \leq X(N)$ for $N \geq 1$. The sequence of matrices $\mathbf{R}(N,1)$ is obviously nondecreasing and tends to R, so that the preceding inequality implies that $R \leq X^*$, and therefore $\mathbf{R} = X^*$.

Before proceeding with the next theorem, another quantity called *Spectral Radius* needs to be defined. The eigenvalue of R with largest absolute value is called the spectral radius of R, denoted $sp(\mathbf{R})$ and provided $\bar{P}$ is irreducible, it is real and positive.

Theorem 2 *If the Markov chain $\bar{P}$ is positive recurrent, then*

(a) *for $i \geq 0$,*

$$p_{i+1} = p_i R \qquad \text{(A.22)}$$

(b) *the eigenvalues of R lie inside the unit disc*

(c) *the matrix B[R] defined as*

$$B[R] = \sum_{k=0}^{\infty} R^k B_k \tag{A.23}$$

is stochastic, and

(d) *the vector p_0 is a positive, left invariant eigenvector of B[R], normalised by*

$$p_0(I - R)^{-1}e = 1 \ . \tag{A.24}$$

Proof: By conditioning on the time and the state of the last visit to the set **i**, if there is such a visit, the relation

$$P^{(n)}_{i+1,j;i+1,j} = {}_{\mathbf{i}}P^{(n)}_{i+1,j;i+1,j} + \sum_{v=1}^{m}\sum_{r=0}^{n} P^{(r)}_{i+1,j;i,v}\ {}_{\mathbf{i}}P^{(n-r)}_{i,v;i+1,j} \quad , \quad n \geq 1 \tag{A.25}$$

is obtained. Adding these equations for n ranging from 1 to N and dividing the resulting sums by N, as $N \to \infty$, and noting that $\sum_{n=1}^{\infty}\ {}_{\mathbf{i}}P^{(n)}_{i+1,j;i+1,j}$ tends to zero, yields:

$$\begin{aligned} P_{i+1,j} &= \lim_{N\to\infty}\{\sum_{v=1}^{m}\frac{1}{N}\sum_{n=1}^{N}\sum_{r=0}^{n} P^{(r)}_{i+1,j;i,v}\ {}_{\mathbf{i}}P^{(n-r)}_{i,v;i+1,j}\} \\ &= \lim_{N\to\infty}\{\sum_{v=1}^{m}\frac{1}{N}\sum_{r=0}^{N} P^{(r)}_{i+1,j;i,v}\sum_{n=0}^{N-r}\ {}_{\mathbf{i}}P^{(n)}_{i,v;i+1,j}\} \\ &= \sum_{v=1}^{m} P_{iv}R_{vj} \end{aligned} \tag{A.26}$$

Note that $N^{-1}\sum_{r=0}^{N} P^{(r)}_{i+1,j;i,v}$ tends to P_{iv} and $\sum_{n=0}^{N}\ {}_{\mathbf{i}}P^{(n)}_{i,v;i+1,j}$ has the limit R_{vj} as $N \to \infty$.

We now consider the expected number of transitions before the first return to the level **0**, given that the chain starts in the state (0,j). This quantity is finite if and only if the chain is positive recurrent. It is given by the j^{th} component of the vector $\sum_{k=1}^{\infty} R^k e$, which is finite if and only if the matrix $\sum_{k=1}^{\infty} R^k = (I - R)^{-1}$ is finite, or equivalently if all the eigenvalues of R lie inside the unit disc, i.e. $sp(R) \leq 1$.

Since the matrix $\bar{P}$ is stochastic, for $k \geq 0$

$$B_k e + \sum_{v=0}^{k} A_v e = e \ . \tag{A.27}$$

Then:

$$\begin{aligned} B[R]e &= \sum_{k=0}^{\infty} R^k B_k e \\ &= (I-R)^{-1}e - \sum_{k=0}^{\infty} R^k \sum_{v=0}^{k} A_v e \\ &= (I-R)^{-1}e - \sum_{v=0}^{\infty}\sum_{k=v}^{\infty} R^k A_v e \\ &= (I-R)^{-1}[I - \sum_{v=0}^{\infty} R^v A_v]e = e \ . \end{aligned} \tag{A.28}$$

Substituting equation (A.22) into equation (A.9) gives:

$$\begin{aligned} p_0 &= \sum_{v=0}^{\infty} p_v B_v \\ &= p_0 B_0 + p_0 R B_1 + p_0 R^2 B_2 + \cdots \\ &= p_0 \sum_{v=0}^{\infty} R^v B_v \\ &= p_0 B[R] \ . \end{aligned} \tag{A.29}$$

Further, the normalising equation $\sum_{k=0}^{\infty} p_k e = 1$ implies:

$$\begin{aligned} \sum_{k=0}^{\infty} p_k e &= p_0 \sum_{k=0}^{\infty} R^k e \\ &= p_0 (I-R)^{-1} e = 1 \ . \end{aligned} \tag{A.30}$$

Hence all parts of the theorem are proved. Note that parts (b) and (d) of this theorem provide the sufficient conditions for the positive recurrence of the Markov chain $\bar{P}$. The following two theorems are stated without proof (see p. 12 of [89] for proofs):

Lemma 4 *If the matrix $\bar{P}$ is irreducible, the matrix R cannot have columns that are identically zero, and R must have at least one positive eigenvalue.*

Theorem 3 *If the Markov chain $\bar{P}$ is irreducible, if the minimal non-negative solution R of the matrix equation (A.13) has all its eigenvalues inside the unit disk, and if the stochastic matrix $B[R]$ has a left positive left invariant vector, then the Markov chain $\bar{P}$ is positive recurrent.*

A.3.1 Complex Boundary Behaviour

The stochastic matrix $\bar{P}$ may be modified to have a more complicated structure near the lower boundary which will lead to *modified matrix geometric invariant vectors* (p. 24 of [89]). Let us assume that matrix $\bar{P}$ may be partitioned as following:

$$\bar{P} = \begin{bmatrix} B_{00} & B_{01} & 0 & 0 & 0 & 0 & \cdots \\ B_{10} & B_{11} & A_0 & 0 & 0 & 0 & \cdots \\ B_{20} & B_{21} & A_1 & A_0 & 0 & 0 & \cdots \\ B_{30} & B_{31} & A_2 & A_1 & A_0 & 0 & \cdots \\ B_{40} & B_{41} & A_3 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{A.31}$$

where the matrices B_{00} and B_{01} are of dimensions $(m_1 - m) \times (m_1 - m)$ and $(m_1 - m) \times m$ respectively. The matrices B_{k0}, $k \geq 1$, are $m \times (m_1 - m)$, while the matrices B_{k1}; $k \geq 1$, A_k; $k \geq 0$, are square matrices of order m. The first m_1; $m_1 > m$, states are called the boundary states.

Assume that matrix $\bar{P}$ is irreducible and $A = \sum_{k=0}^{\infty} A_k$ is stochastic. The probability vector p is partitioned into an $(m_1 - m)$-vector p_0 and m-vectors $p_k, k \geq 1$. The matrix R is defined as before. The matrix $B[R]$ is defined as

$$B[R] = \begin{bmatrix} B_{00} & B_{01} \\ \sum_{k=1}^{\infty} R^{k-1} B_{k0} & \sum_{k=1}^{\infty} R^{k-1} B_{k1} \end{bmatrix} . \tag{A.32}$$

Lemma 5 *If $sp(R) < 1$, the matrix $B[R]$ is stochastic.*

Proof : It suffices to verify that

$$\sum_{k=1}^{\infty} R^{k-1}(B_{k0} + B_{k1})e = \sum_{k=1}^{\infty} R^{k-1}(I - \sum_{v=0}^{k-1} A_v)e = e \ .$$

This is done by direct calculation as in equation (A.28).

Theorem 4 *The irreducible Markov chain $\bar{P}$ is positive recurrent if and only if $sp(R) < 1$ and the stochastic matrix $B[R]$ has a positive left invariant m_1-vector $[p_0, p_1]$. Normalising the vector $[p_0, p_1]$ as*

$$p_0 e + p_1 (I - R)^{-1} e = 1 \tag{A.33}$$

the invariant probability vector p of $\bar{P}$ is given for $k \geq 1$, by

$$p_k = p_1 R^{k-1} \tag{A.34}$$

Proof : It suffices to verify that the stated vector p satisfies the equations

$$p = p\bar{P}$$

$$pe = 1 \tag{A.35}$$

and this straightforward.

Remarks :

- The matrix $\bar{P}$ in (A.31) differs from the canonical form of (A.5) only in the definition of the transition probabilities from the boundary states. The probabilistic significance of the matrix R therefore remains the same, but applies only to the non-boundary states.

- The vector $p = [p_0, p_1, p_1 R, p_1 R^2, \cdots]$ is known as the modified matrix-geometric invariant vector of $\bar{P}$. The vector p_0 may be further partitioned for the ease of computation.

- Noting that in the matrix $\bar{P}$ of (A.31), the submatrix $[B_{00}, B_{01}]$, by which the first m_1 columns protrude above the regular pattern formed by the other columns, must have fewer rows than columns, otherwise the Markov chain $\bar{P}$ would be reducible.

- There are some forms of partitioned matrices that are superficially similar to the form of $\bar{P}$ in (A.31), but that need to be appropriately repartitioned before this methods of analysis can be applied to them.

A.3.2 Continuous Parameter Markov Processes

There is an entirely analogous theory for continuous Markov processes whose infinitesimal generator Q may be partitioned similarly to the matrix $\bar{P}$. Here, a continuous-parameter Markov process is considered with the generator Q of the form

$$Q = \begin{bmatrix} B_0 & A_0 & 0 & 0 & 0 & \dots \\ B_1 & A_1 & A_0 & 0 & 0 & \dots \\ B_2 & A_2 & A_1 & A_0 & 0 & \dots \\ B_3 & A_3 & A_2 & A_1 & A_0 & \dots \\ B_4 & A_4 & A_3 & A_2 & A_1 & \dots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} . \tag{A.36}$$

The off-diagonal elements are non-negative. The diagonal elements are all strictly negative, and

$$\sum_{v=0}^{k} A_v e + B_k e = 0 \quad , \text{ for } k \geq 0 \ . \tag{A.37}$$

Positive recurrence of Q is equivalent to the existence of a positive probability vector $p = [p_0, p_1, p_2, \dots]$ which satisfies the equations

$$\sum_{v=0}^{\infty} p_v B_v = 0$$

$$\sum_{v=0}^{\infty} p_{k+v-1} A_v = 0 \quad , \text{ for } k \geq 1 \ . \tag{A.38}$$

Theorem 5 *The irreducible Markov process Q is positive recurrent if and only if the minimal non-negative solution R of the equation*

$$\sum_{k=0}^{\infty} R^k A_k = 0 \tag{A.39}$$

has $sp(R) < 1$, and if there exists a positive vector p_0 such that

$$p_0 B[R] = p_0 \sum_{k=0}^{\infty} R^k B_k = 0 \ . \tag{A.40}$$

The matrix $B[R] = \sum_{k=0}^{\infty} R^k B_k$ *is a generator. The stationary probability vector* p, *satisfying* $pQ = \mathbf{0}$, $pE = 1$, *is given by*

$$p_k = p_0 R^k \quad , \text{ for } k \geq 0 \tag{A.41}$$

and p_0 *is normalised by*

$$p_0 (I - R)^{-1} e = 1 \ . \tag{A.42}$$

The matrix R *satisfies* $sp(R) < 1$ *if and only if*

$$\pi A_0 e < \sum_{k=2}^{\infty} (k-1) \pi A_k e \tag{A.43}$$

where π *is given by* $\pi A = \mathbf{0}$ *and* $\pi e = 1$.

Proof : Let τ be any real number satisfying

$$\tau \geq \max_j \{\max[-(B_0)_{jj}, -(A_1)_{jj}]\} > 0 \ . \tag{A.44}$$

Then (A.38) can be rewritten as

$$\begin{aligned} p_0 &= \sum_{v=0}^{\infty} p_v B'_v \\ p_k &= \sum_{v=0}^{\infty} p_{k+v-1} A'_v \quad , \text{for} k \geq 1 \end{aligned} \tag{A.45}$$

where $B'_v = \delta_{v0} I + \tau^{-1} B_v$, and $A'_v = \delta_{v1} I + \tau^{-1} A_v$, for $v \geq 0$, and where δ is the Kronecker delta function. Since (A.45) is exactly similar to (A.9), the recurrence properties of the process Q can therefore be deduced from those of the chain $\bar{P}'$ with matrix of the form

$$\bar{P}' = \begin{bmatrix} B'_0 & A'_0 & 0 & 0 & 0 & \ldots \\ B'_1 & A'_1 & A'_0 & 0 & 0 & \ldots \\ B'_2 & A'_2 & A'_1 & A'_0 & 0 & \ldots \\ B'_3 & A'_3 & A'_2 & A'_1 & A'_0 & \ldots \\ B'_4 & A'_4 & A'_3 & A'_2 & A'_1 & \ldots \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{bmatrix} \tag{A.46}$$

and the statements of this theorem follow immediately from the results for the discrete case, applied to the matrix $\bar{P}'$. It suffices then to replace the matrices A'_v and B'_v by their definition in terms of A_v and B_v, for $v \geq 0$. Other minor details of this theorem are verified in page 33 of [89].

A.3.3 Quasi-Birth-and-Death (QBD) Processes

These processes are particular cases of Markov processes where the infinitesimal generator matrix is *block tridiagonal*, i.e. a quasi-birth-death process (p. 82 of [89]) is a Markov process on the state space $E = \{(i,j),\ i \geq 0,\ 1 \leq j \leq m\}$, with infinitesimal generator Q, given by

$$Q = \begin{bmatrix} B_0 & A_0 & 0 & 0 & 0 & \dots \\ B_1 & A_1 & A_0 & 0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & 0 & A_2 & A_1 & \dots \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{bmatrix} \tag{A.47}$$

where

$$B_0 e + A_0 e = B_1 e + A_1 e + A_0 e = (A_0 + A_1 + A_2)e = \mathbf{0} \ .$$

The matrix Q is also assumed to be irreducible. The matrix $A = A_0 + A_1 + A_2$ is a finite generator. The following theorem is a particular case of Theorem 5.

Theorem 6 *The process Q is positive recurrent if and only if the minimal non-negative solution R to the matrix-quadratic equation*

$$R^2 A_2 + R A_1 + A_0 = 0 \tag{A.48}$$

has all its eigenvalues inside the unit disc, and the finite system of equations

$$\begin{aligned} p_0(B_0 + R B_1) &= 0 \\ p_0(I - R)^{-1} e &= 1 \end{aligned} \tag{A.49}$$

has a unique positive solution p_0.

If the matrix A is irreducible, then $sp(R) < 1$ if and only if

$$\pi A_2 e > \pi A_0 e \tag{A.50}$$

where π is the stationary probability vector of A. The stationary probability vector $p = [p_0, p_1, p_2, ...]$ of Q is given by

$$p_i = p_0 R^i \quad , \text{ for } i \geq 0 \ . \tag{A.51}$$

The (equivalent) qualities

$$RA_2e - A_0e = RB_1e - B_0e = \mathbf{0}$$

hold.

Appendix B

Brief Correlation Theory

The purpose of this appendix is to give a brief description of correlation analysis to the extent that it has been used in chapter 6. Let X and Y be two random variables. The discrete correlation between X and Y with shift m is given by $R_{XY}(m)$ defined as

$$R_{XY}(m) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} X(i)Y(i+m)$$

Let $\bar{x}$ and $\bar{y}$ be the average values of $X(i)$ and $Y(i)$, i.e.

$$\begin{aligned} \bar{x} &= \frac{1}{N} \sum_{i=0}^{N-1} X(i) \\ \bar{y} &= \frac{1}{N} \sum_{i=0}^{N-1} Y(i) \end{aligned}$$

We can rewrite $X(i)$ and $Y(i)$ as

$$\begin{aligned} X(i) &= x(i) + \bar{x} \\ Y(i) &= y(i) + \bar{y} \end{aligned}$$

where $x(i)$ and $y(i)$ are the autocorrelation equivalents forms of $X(i)$ and $Y(i)$. It is obvious that

$$\begin{aligned} \frac{1}{N} \sum_{i=0}^{N-1} x(i) &= 0 \\ \frac{1}{N} \sum_{i=0}^{N-1} y(i) &= 0 \end{aligned}$$

Hence $R_{XY}(m)$ may be rewritten as

$$\begin{aligned} R_{XY}(m) &= \frac{1}{N}\sum_{i=0}^{N-1}\{(x(i)+\bar{x})(y(i+m)+\bar{y})\} \\ &= \frac{1}{N}\sum_{i=0}^{N-1}\{x(i)y(i+m)+x(i)\bar{y}+y(i+m)\bar{x}+\bar{x}\bar{y}\} \\ &= \frac{1}{N}\{\sum_{i=0}^{N-1}x(i)y(i+m)+\bar{y}\sum_{i=0}^{N-1}x(i)+\bar{x}\sum_{i=0}^{N-1}y(i+m)\}+\bar{x}\bar{y} \\ &= \frac{1}{N}\sum_{i=0}^{N-1}x(i)y(i+m)+\bar{x}\bar{y} \\ &= R_{xy}(m)+\bar{x}\bar{y} \end{aligned}$$

Therefore

$$R_{xy}(m) = R_{XY}(m) - \bar{x}\bar{y}$$

Or in the normalised form :

$$\begin{aligned} \hat{R}_{xy}(m) &= \frac{R_{xy}(m)}{R_{xy}(0)} \\ &= \frac{R_{XY}(m)-\bar{x}\bar{y}}{R_{XY}(0)-\bar{x}\bar{y}} \end{aligned}$$

References

[1] E. Larsen. *'Telecommunications: a history'.* Frederick Muller Limited, London, 1977. page 31.

[2] M. De Prycker, R. Peschi and T. Van Landegem. 'B-ISDN and OSI protocol reference model'. *IEEE Network*, 7(2):10–18, March 1993.

[3] I. M. Leslie, D. R. McAuley and D. L. Tennenhouse. 'ATM Everywhere?'. *IEEE Network*, 7(2):40–46, March 1993.

[4] W. R. Byrne, G. W. R. Luderer, G. Clapp, B. L. Nelson and H. J. Kafka. 'Evolution of metropolitan area networks to Broadband ISDN'. *IEEE Communications Magazine*, pages 69–82, January 1991.

[5] L. Kleinrock. 'ISDN - The path to broadband networks'. *Proceedings of the IEEE*, 79(2):112–117, February 1991.

[6] W. Zitsen. 'Metropolitan area networks: taking LANs into the public network'. *Telecommunications*, pages 53–60, June 1990.

[7] A. Forcina A. Biocca, G. Freschi and R. Melen. 'Architectural issues in the interoperability between MANs and the ATM network'. In *Proceedings of the XIII International Switching Symposium*, volume 2, pages 23–28, Stockholm, 1990.

[8] J.F. Mollenauer. 'Standards of metropolitan area networks'. *IEEE Communications Magazine*, 26(4):15–19, April 1988.

[9] ISO. *'ISO 9314-1,-2,-3: Fibre Distributed Data Interface (FDDI)'.* American National Standards Association, New York.

[10] F. Ross. 'An overview of FDDI: the Fiber Distributed Data Interface'. *IEEE Journal on Selected Areas in Communications*, 7(7), September 1989.

[11] ANSI. *'Hybrid ring control'*. May 1990. Revision 6.

[12] K. Caves. 'FDDI-II : A new standard for integrated services high speed LANs'. *Telecommunications*, September 1987.

[13] IEEE. *'IEEE Std 802.6 : Distributed Queue Dual Bus (DQDB) subnetwork of a Metropolitan Area Network (MAN)'*. 1991.

[14] CCITT. *'Recommendation I.120 : Integrated Services Digital Networks (ISDNs)'*. Geneva, 1988.

[15] Martin De Prycker. 'Evolution form ISDN to BISDN: a logical step towards ATM'. *Computer Communications*, 12(3):141–146, June 1989.

[16] CCITT. *'Recommendation I.113 : Vocabulary of terms for broadband aspects of ISDN'*. International Telecommunications Union, Geneva, 1991.

[17] CCITT. *'Recommendation I.121 : Broadband aspects of ISDN'*. International Telecommunications Union, Geneva, 1991.

[18] CCITT. *'Recommendation I.211 : B-ISDN service aspects'*. International Telecommunications Union, Geneva, 1991.

[19] CCITT. *'Recommendation I.320 : ISDN Protocol Reference Model'*. International Telecommunications Union, Geneva, 1989.

[20] CCITT. *'Recommendation I.321 : B-ISDN Protocol Reference Model and its application'*. International Telecommunications Union, Geneva, 1991.

[21] CCITT. *'Recommendation X.200 : Reference model of open systems interconnection for CCITT applications'*. International Telecommunications Union, Melbourne, 1988.

[22] CCITT. *'Recommendation I.432 : B-ISDN user-network interface - physical layer specification'*. International Telecommunications Union, Geneva, 1991.

[23] CCITT. *'Recommendation I.150 : B-ISDN ATM functional characteristics'.* International Telecommunications Union, Geneva, 1991.

[24] CCITT. *'Recommendation I.311 : B-ISDN general network aspects'.* International Telecommunications Union, Geneva, 1991.

[25] John Burgin and Dennis Dorman. 'Broadband ISDN resource management: the role of Virtual Paths'. *IEEE Communications Magazine*, pages 44–48, September 1991.

[26] CCITT. *'Recommendation I.361 : B-ISDN ATM layer specifications'.* International Telecommunications Union, Geneva, 1991.

[27] G.M. Woodruff, R.G.H. Rogers and P.S. Richards. 'A congestion control framework for high-speed integrated packetized transport'. In *Proceedings of IEEE Global Telecommunications Conference*, pages 203–207, Florida, November 1988.

[28] G.M. Woodruff and R. Kositpaiboon. 'Multimedia traffic management principles for guaranteed ATM network performance '. *IEEE Journal on Selected Areas in Communications*, 8:437–446, 1990.

[29] K. Kawashima and H. Saito. 'Teletraffic issues in ATM networks'. In *'ITC Specialist Seminar'*, Adelaide, 1989.

[30] A.E. Eckberg, D.T. Luan and D.M. Lucantoni. 'Meeting the challenge: congestion and flow control strategies for broadband information transport'. In *Proceedings of IEEE Global Telecommunications Conference*, pages 49.3.1–49.3.5, Dallas-Texas, November 1989.

[31] C. J. May R. Dighe and G. Ramamurthy. 'Congestion avoidance strategies in broadband packet networks'. In *Proceedings of IEEE Infocom*, volume 1, pages 295–303, Bal Harbour, USA, April 1991.

[32] CCITT. *'Draft Recommendation I.371 : Traffic control and resource management in B-ISDN'.* Melbourne, 1991.

[33] M. Zukerman and S. Chan. 'Fairness in Broadband ISDN'. In *The Proceedings of IEEE Infocom '92*, pages 2241–2250, Florence-Italy, May 1992.

[34] J. W. Roberts. 'Traffic control in the B-ISDN'. *Computer Networks and ISDN Systems*, 25(10):1055–1064, May 1993.

[35] G. Rigolio G. Gallassi and L. Fratta. 'ATM: bandwidth assignment and bandwidth enforcement policies'. In *Proceedings of IEEE Global Telecommunications Conference*, pages 49.6.1–49.6.6, Dallas-Texas, November 1989.

[36] L. Dittmann and S. B. Jacobsen. 'Statistical multiplexing of identical bursty sources in an ATM network'. In *Proceedings of IEEE Global Telecommunications Conference*, pages 39.6.1–39.6.5, Florida, November 1988.

[37] S. Akhtar. 'congestion control in a fast packet switching network'. Master's thesis, Washington University, St. Louis, 1987.

[38] Xiaoqiang Chen and I. M. Leslie. 'Performance evaluation of input traffic Ccntrol'. In *Proceedings of IEEE Infocom*, pages 552–561, Florence-Italy, May 1992.

[39] A. E. Eckberg. 'Generalised peakedness of teletraffic processes'. In *Proceedings of the 10th International Teletraffic Congress*, Montreal, 1983.

[40] J. Y. Hui and E. Arthurs. 'A broadband packet switch for integrated transport'. *IEEE Journal on Selected Areas in Communications*, 5:1264–1273, October 1987.

[41] K. Sriram and W. Whitt. 'Characterising Superposition Arrival Processes in Packet Multiplexers for Voice and Data'. *IEEE Journal on Selected Areas in Communications*, 4:833–846, September 1986.

[42] M. Kawakatsu K. Nakamaki and A. Notoya. 'Traffic control for ATM networks'. In *Proceedings of IEEE ICC*, pages 22.5.1–22.5.5, 1989.

[43] M. Hirano and N. Watanabe. 'Characteristics of a cell multiplexer for bursty ATM traffic'. In *Proceedings of IEEE ICC*, pages 13.2.1–13.2.5, 1989.

[44] T. Kamitake and T. Suda. 'Evaluation of an admission control scheme for an ATM network considering fluctuations in the cell loss rate'. In *Proceedings of*

IEEE Global Telecommunications Conference, pages 49.4.1–49.4.7, Dallas-Texas, November 1989.

[45] S. Q. Li. 'Study of information loss in packet voice systems'. *IEEE Transactions on Communications*, 37:1192–1202, November 1989.

[46] E. P. Rathgeb. 'Modelling and performance comparison of policing mechanisms for ATM networks'. *IEEE Journal on Selected Areas in Communications*, 9(3):325–334, April 1991.

[47] J. S. Turner . 'New directions in communications (or which way to the information age?) '. *IEEE Communications Magazine*, 25:8–15, 1986.

[48] L. Dittmann, S. B. Jacobsen and K. Moth. 'Flow enforcement algorithms for ATM networks'. *IEEE Journal on Selected Areas in Communications*, 9(3):343–350, April 1991.

[49] I. Cidon and I.S. Gopal. 'PARIS : An approach to integrated high speed private networks '. *International Journal of Digital & Analog Cabled Systems*, 1:77–86, 1988.

[50] A.E. Eckberg, D.T. Luan and D.M. Lucantoni. 'Bandwidth management : a congestion control strategy for broadband packet networks - characterizing the throughput-burstiness filter '. In *ITC Specialist Seminar*, Adelaide, September 1989.

[51] M. Butto, E. Cavallero and A. Tonietti. 'Effectiveness of the "Leaky Bucket" policing mechanism in ATM networks'. *IEEE Journal on Selected Areas in Communications*, 9(3):335–342, April 1991.

[52] Young Han Kim. 'Performance analysis of leaky-bucket bandwidth enforcement strategy for bursty traffics in an ATM network'. *Computer Networks and ISDN Systems*, 25(3):295–303, September 1992.

[53] H. Kröner. 'Comparative performance study of space priority mechanisms for ATM Networks'. In *Proceedings of IEEE Infocom'90*, pages 1136–1143, san Francisco, 1990.

[54] F. Bonomi, L. Fratta, S. Montagna and R. Paglino. 'Priority on cell service and on cell loss in ATM switching'. In *Proceedings of the 7th ITC Seminar*, Morristown, NJ, October 1990. paper 7.2.

[55] T. C. Hou and A. K. Wong. 'Queueing analysis for ATM switching of mixed continuous-bit-rate and bursty traffic'. In *Proceedings of Infocom '90*, pages 660–667, San Francisco, CA, June 1990.

[56] N. Yin, S. Q. Li and T. E. Stern. 'Congestion control for packet voice by selective packet discarding'. *IEEE Transaction on Communications*, 38(5):674–683, May 1990.

[57] J. F. Chang and C. S. Wu. 'The effect of prioritization of a concentrator under an accept, otherwise reject strategy'. *IEEE Transactions on Communications*, 38(7):1031–1039, July 1990.

[58] H. Kröner, G. Hebuterne, P. Boyer and A. Gravey. 'Priority management in ATM switching nodes'. *IEEE Journal on Selected Areas in Communications*, 9(7):418–427, April 1991.

[59] D. M. Lucantoni and S. P. Parekh. 'Selective cell discard mechanisms for a B-ISDN congestion control architecture'. In *Proceedings of the 7th ITC Seminar*, Morristown, NJ, October 1990. paper 10.3.

[60] C. J. O'Neill. 'Fairness discarding for congestion control in ATM networks'. In *Proceedings of the Australian Broadband Switching and Services Symposim '92*, volume 1, pages 185–192, Melbourne, July 1992.

[61] C. J. O'Neill. 'A method for congestion control in ATM networks using peak rate throttling'. *International Journal of Digital and Analogue Communication Systems*, 4:1–10, 1991.

[62] J. B. Nagle. 'On packet switches with infinite storage'. *IEEE Transactions on Communications*, 35(4):435–438, April 1987.

[63] X. Chen and I. M. Leslie. 'Neural adaptive congestion control for broadband ATM networks'. *IEE Proceedings-I*, 139(3):233–240, June 1992.

[64] H. E. Rauch and T. Winarske. 'Neural networks for routing communication traffic'. *IEEE Control Systems Magazine*, (4):26–31, 1988.

[65] A. Hiramatsu. 'ATM communications network control by neural networks'. *IEEE Transactions on Neural Networks*, 1(1):122–130, 1990.

[66] M. K. A. Mehmet and F. Kamoun. 'A neural network approach to the maximum flow problem'. In *Proceedings of IEEE Globecom '91*, pages 130–134, Phoenix-Arizona, 1991.

[67] L. Kleinrock. *'Queueing systems, volume II: computer applications'*. John Wiley & Sons, 1975.

[68] H. Akimaru and H. Takahashi. 'An approximate formula for estimating individual call losses in overflow systems'. *IEEE Transactions on Communications*, 31(6):808–811, 1983.

[69] J. Matsumoto and Y. Watanabe. 'Individual traffic characteristics of queueing systems with multiple Poisson and overflow inputs'. *IEEE Transactions on Communications*, 33(1):1–9, 1985.

[70] L. A. Gimpelson. 'Analysis of mixtures of wide- and narrow-band traffic '. *IEEE Transactions on Commununication Technology*, 13(3):258–266, 1965.

[71] U. N. Bhat and M.J. Fischer. 'Multichannel queueing systems with heterogeneous classes of arrivals'. *Naval Research Logistics Quarterly*, 23(1):271–282, 1976.

[72] B. Kraimeche. *'Traffic access control strategies in integrated service digital networks '*. PhD thesis, Columbia University, 1984.

[73] M.J. Ross and O.A. Mowafi. 'Performance analysis of hybrid switching concepts for integrated voice/data communications '. *IEEE Transactions on Communications*, 30(5):1073–1087, 1982.

[74] K. Kummerle. ' Multiplex performance for integrated line and packet-switched traffic'. In *Proceedings of International Conf. Comput. Commun.*, pages 517–523, 1974.

[75] G.J. Coviello and P.A. Vena. Integration of circuit/packet switching by a SENET (Slotted Envelope NETwork) concept. In *Proceedings of National Telecommunications Conference*, pages 42.12–42.17, 1975.

[76] M.J. Fischer and T.C.Harris. 'A model for evaluating the performance of an integrated circuit- and packet-switched multiplex structure'. *IEEE Transactions on Communications*, 24(2):195–202, 1976.

[77] T. Yamaguchi and M. Akiyama. 'An integrated hybrid traffic switching system mixing preemptive wideband and waitable narrowband calls'. *Elect. and Commun. in Japan*, 53(5):43–52, 1970.

[78] M.J.Fischer. 'A queueing analysis of an integrated telecommunications system with priorities'. *INFOR*, 15(3):277–288, 1977.

[79] A. J. Bialley, A. J. McLaughlin and C. J. Weinstein. 'Voice communication in integrated digital voice and data networks'. *IEEE Transactions on Communications*, 28(9):1478–1490, 1980.

[80] C.J. Weinstein et al. 'Data traffic performance of an integrated circuit- and packet- switched multiplex structure'. *IEEE Transactions on Communications*, 22(6):873–878, 1980.

[81] R.M. Feldman and C.A. Claybaugh. 'A note on a computational model for a data/voice communication queuing system'. *Naval Research Logistics Quarterly*, 29(5):529–534, 1982.

[82] D.P. Gaver and J.P. Lehoczky. 'Channels that cooperatively service a data stream and voice messages'. *IEEE Transactions on Communications*, 30(5):1153–1162, 1982.

[83] K. Sriram, P. K. Varshney and J. G. Shanthikumar. 'Discrete-time analysis of integrated voice/data multiplexers with and without speech activity detectors'. *IEEE Journal on Selected Areas in Communications*, 1(6):1124–1132, 1983.

[84] Y. D. Serres and L.G. Mason. 'A Multiserver queue with narrow-band and wide-band customers and WB restricted access'. *IEEE Transactions on Communications*, 36(6):675–684, 1988.

[85] K.C. Chua. *Mixed voice/data packet switching and traffic access control strategies in B-ISDN.* PhD thesis, Dept. of Electrical and Electronic Engineering, University of Auckland, 1990.

[86] D.J.H. Lewis and C. Ambepitiya. 'Queueing delays on subscriber links carrying voice and data traffic '. In *ATERB Fast Packet Switching Workshop*, OTC, Sydney, 1989.

[87] Moshe Zukerman and Paul Kirton. 'Queueing analysis of a B-ISDN switching system'. In *Australian Fast Packet Switching Workshop*, Melbourne, 1988.

[88] Moshe Zukerman. 'Applications of matrix-geometric solutions for queueing performance evaluation of a hybrid switching system'. *Journal of Australian Mathematical Society, Ser B 31*, pages 219–239, 1989.

[89] M.F. Neuts. *'Matrix-geometric solutions in stochastic models'.* Johns Hopkins University Press, 1981.

[90] A. Papoulis. *'Probability, random variables and stochastic processes'.* McGraw Hill, 1984.

[91] M.F. Neuts. 'Further results on the M/M/1 queue with randomly varying rates'. *OPSEARCH*, 15(4):158–168, 1978.

[92] M.F. Neuts and D.M. Lucantoni. 'A Markovian queue with N servers subject to breakdowns and repairs'. *Management Science*, 25(9):849–861, 1979.

[93] Kishor Shridharbhai Trivedi. *'Probability & statistics with reliability, queueing and conputer science applications'.* Prentice-Hall, 1982.

[94] M.H. Rossiter. 'A switched Poisson model for data traffic'. *Australian Telecommunications Research*, 21:1–11, 1987.

[95] B. Kraimeche and M. Schwartz. 'Analysis of traffic access control strategies in integrated services digital networks'. *IEEE Transactions on Communications*, 33(10):1085–1093, 1985.

[96] Daryoush Habibi, D.J.H. Lewis and D.T. Nguyen. 'Access control in ATM networks carrying video, interactive images and data traffic'. In *Australian Broadband Switching and Services Symposium*, pages 165–174, Sydney, July 1991.

[97] D.L. McLaren and D.T. Nguyen. 'An ATM-compatible video coding scheme using psychovisual compression'. In *'ICCS 90'*, volume 2, Singapore, 1990.

[98] A.W. Johnson and A. Tessarolo. 'Timing recovery for video codecs operating on broadband packet based networks '. In *Australian Video Communications Workshop*, pages 146–155, Melbourne, July 1990.

[99] David L. McLaren. 'Access vontrol of ATM packet video : a question of priority'. In *Australian Broadband Switching and Services Symposium*, volume 2, pages 174–181, Sydney, July 1991.

[100] CCITT. *'Recommendation I.362 : B-ISDN ATM Adaption Layer (AAL) Functional Description'*. International Telecommunications Union, Geneva, 1991.

[101] Daryoush Habibi, DJH Lewis, DT Nguyen and J Pieloor. 'Analysis of an access node multiplexer in a system serving CBR & VBR traffic'. To Appear in: *Computer Communications(Butterworth/Heinemann)*, 16(12), 1993.

[102] Daryoush Habibi, D.J.H. Lewis, D.T. Nguyen and J. Pieloor. 'Performance of a multiplexer in a B-ISDN network with STM and ATM traffic'. In *Australian Broadband Switching and Services Symposium*, pages 691–698, Melbourne, July 1992.

[103] D.J.H. Lewis and Daryoush Habibi. 'Analysis v. simulation: the computational effort '. In *Australian Broadband Switching and Services Symposium*, volume 1, page 275, Sydney, July 1991.

[104] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J.D. Robbins. 'Performance models of statistical multiplexing in packet video communications'. *IEEE Transactions on Communications*, 36(7):834–844, 1988.

[105] M. Nomura, T. Fujii and N. Ohta. 'Basic characteristics of variable rate video coding in ATM environment'. *IEEE Journal on Selected Areas in Communications*, 7(5):752–760, June 1989.

[106] Kishimoto, Ogata and Inumaru. 'Generation interval distribution characteristics of packetized variable rate video coding data streams in an ATM network'. *IEEE Journal on Selected Areas in Communications*, 7(5):833–841, June 1989.

[107] S. S. Dixit and P. Skelly. 'Video traffic smoothing and ATM multiplexer performance'. In *Proceedings of IEEE Globecom '91*, Phoenix-Arizona, 1991.

[108] P. Sen, B. Maglaris, N. Rikli and D. Anastassiou. 'Models for packet switching of variable-bit-rate video sources '. *IEEE Journal on Selected Areas in Communications*, 7(5):865–869, 1989.

[109] L. Kleinrock. *'Queueing systems, volume I: theory'*. John Wiley & Sons, 1975.

[110] Ferit Yegenoglu, Bijan Jabbari and Ya-Qin Zhang. 'Modelling of motion classified VBR video codecs'. In *IEEE INFOCOM Proceedings*, volume 1, pages 105–109, Florence, Italy, May 1992.

[111] H. Yamada, K. Miyake, F. Kishino and K. Manabe. 'Modeling of arrival process of packetized video and related statistical multiplexer performance'. In *Proceedings of IECEJ National Conference*, 1989.

[112] Y. Yasuda, H. Yasuda, N. Ohta and F. Kishino. 'Packet video transmission through ATM networks'. In *Proceedings of IEEE Global Telecommunications Conference*, pages 25.1.1–25.1.5, Dallas-Texas, November 1989.

[113] B. Sengupta B. Melamed, D. Raychaudhuri and J. Zdepski. 'TES-based traffic modelling for performance rvaluation of integrated networks'. In *Proceedings of IEEE Infocom '92*, pages 75–84, Florence-Italy, May 1992.

[114] Paul Skelly, Sudhir Dixit and Mischa Schwartz. 'A histogram-based model for video traffic behaviour in an ATM network node with an application to congestion control'. In *IEEE INFOCOM Proceedings*, volume 1, pages 95–103, Florence, Italy, May 1992.

[115] B. Melamed. 'TES: a class of methods for generating autocorrelated uniform variates'. In *NEC Research Institute, INC.*, Princeton, New Jersey, 1990.

[116] D. L. Jagerman and B. Melamed. 'The autocovariance structure of TES processes'. In *NEC Research Institute, INC.*, Princeton, New Jersey, 1990.

[117] R. W. Wolff. *'Stochastic modelling and the theory of queues'*. Prentice Hall, New Jersey, 1989.

[118] Daryoush Habibi. 'A hidden Markov model for modelling VBR video traffic'. In *The Seventh Australian Teletraffic Research Seminar*, pages 181–190, River Murray, South Australia, November 1992.

[119] L. R. Rabiner. 'A tutorial on hidden Markov models and selected applications in speech recognition'. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[120] L. E. Baum and T. Petrie. 'Statistical inference for probabilistic functions if finite state Markov chains'. *The Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[121] L. E. Baum and G. R. Sell. 'Growth functions for transformations on manifolds'. *Pacific Journal of Mathematics*, 27(2):211–227, 1968.

[122] L. E. Baum, T. Petrie, G. Soules and N. Weiss. 'A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains'. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[123] F. Jelinek, L. R. Bahl and R. L. Mercer. 'Design of a linguistic statistical decoder for the recognition of continuous speech'. *IEEE Transactions on Information Theory*, 21:250–256, 1975.

[124] J. K. Baker. 'The dragon system - an overview'. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29, 1975.

[125] L. R. Bahl and F. Jelinek. 'Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition'. *IEEE Transactions on Information Theory*, 21:404–411, 1975.

[126] F. Jelinek. 'Continuous speech recognition by statistical methods'. *Proceedigs of IEEE*, 64:532–536, April 1976.

[127] B. H. Juang and L. R. Rabiner. 'Mixture autoregressive hidden Markov models for speech signals'. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413, December 1985.

[128] David McLaren . *'Video and image coding for broadband integrated services digital networks'.* PhD thesis, Department of Electrical & Electronic Engineering, University of Tasmania, 1992.

[129] Daryoush Habibi and D.J.H. Lewis. 'Queues with periodic input and output rates'. In *Australian Broadband Switching and Services Symposium*, Wollongong, July 1993.

[130] Daryoush Habibi and D.J.H Lewis. 'A solution for cyclo-stationary queueing systems'. *Submitted for Publication in IEE Electronics Letters.*

[131] G. Latouche. 'Sample path analysis of packet queues subject to periodic traffic'. *Computer Networks and ISDN Systems*, 20:409–413, 1990.

[132] G. Latouche. 'a study of deterministic cycles in packet queues subject to periodic traffic'. Université Libre de Bruxelles, Report 89.1, 1989.

[133] Tedijanto. 'Exact results for the cyclic-service queue with a Bernouli schedule'. *Performance Evaluation*, 11:107–115, 1990.

[134] L.D. Servi. 'Average delay approximation of M/G/1 cyclic service queues with Bernouli schedules'. *IEEE Journal on Selected Areas in Communications*, 4:813–822, 1986.

[135] O. J. Boxma and B. Meister. 'Waiting-time approximation for cyclic-service systems with switchover times'. *Performance Evaluation*, 7:299–308, 1987.

[136] M.J. Ferguson and Y.J. Aminetzah. 'Exact results for nonsymmetric token ring systems'. *IEEE Transactions on Communications*, 33:223–231, 1985.

[137] O.C. Ibe and X. Cheng. 'Approximate analysis of asymmetric single-service token-passing systems'. *IEEE Transactions on Communications*, 37:572–577, 1989.

[138] S.W. Fuhrmann and Y.T. Wang. 'Analysis of cyclic service systems with limited service: Bounds and approximations'. *Performance Evaluation*, 9:35–54, 1988.

[139] K.K. Leung. 'Cyclic-service systems with probabilistically-limited service'. *IEEE Journal on Selected Areas in Communications*, 9(2):185–193, February 1991.

[140] H. Levy and L. Kleinrock. 'Polling systems with zero switch-over periods: a general method for analysing the expected delay'. *Performance Evaluation*, 13:97–107, 1991.

[141] A. Papoulis. *'Signal Analysis'*. McGraw-Hill, 1977.

[142] Duke Hong and Tatsuya Suda. 'Congestion control and prevention in ATM networks'. *IEEE Network*, 5(4):10–16, 1991.

[143] K. Y. Eng, R. D. Gitlin and M. J. Karol. 'A framework for a national broadband (ATM/B-ISDN) network'. In *Proceedings of IEEE Global Telecommunications Conference*, pages 515–519, 1990.

[144] E. Dutkiewicz and G. Anido. 'Traffic management and control in broadband networks'. In *Australian Fast Packet Switching Workshop*, pages 105–112, Melbourne, November 1990.

[145] K.C. Chua. 'Performance of common bank rate control throttle in ATM networks'. *Electronic Letters*, 27(21):1905–1907, 1991.

[146] K. C. Chua and D. T. Nguyen. 'Bit-rate compression and restricted access strategy for integrated services digital networks'. *Computer Communications*, 13(2):67–72, 1990.

[147] Tutomu Murase, Hiroshi Suzuki, Shohei Sato and Takao Takeuchi. 'A call admission control scheme for ATM networks using a simple quality estimate'. *IEEE Journal on Selected Areas in Communications*, 9(9):1461–1470, Dec. 1991.

[148] V. Ramaswami and W. Willinger. 'Efficient traffic performance strategies for packet multiplexers '. In *ITC Specialist Seminar*, pages 4.2.1–4.2.7, Adelaide, 1989.

[149] O. Gihr and P. Tran-Gia. 'A layered description of ATM cell traffic streams and correlation analysis'. *Australian Telecommunication Rerearch*, 24(2):9–17, 1990.

[150] K. W. Fendick, D. Mitra, I. Mitrani, M. A. Rodrigues, J. B. Seery and A. Weiss. 'An approach to high-performance, high-speed data networks'. *IEEE Communications Magazine*, 30(10):74–82, October 1991.

[151] J. W. Roberts. 'Variable-bit-rate traffic control in B-ISDN'. *IEEE Communications Magazine*, pages 50–56, September 1991.

[152] S. Q. Li. 'Generating function approach for discrete queueing analysis with decomposable arrival and service Markov chains'. In *Proceedings of the IEEE Infocom*, pages 2168–2177, Florence-Italy, May 1992.

[153] A. O. Allen. *'Probability, statistics, and queueing theory with computer science applications'.* Academic Press, New York, 1978.

[154] William Feller. *'An introduction to probability theory and its applications'.* John Wiley & Sons, New York, 1950.

[155] Bharucha-Reid, A.T. *'Elements of the theory of Markov processes and their applications'.* McGraw-Hill, 1960.

Reprints of Selected Papers

This section contains reprints of a selection of papers which have been published by the author during the course of this research. The complete list of publications is provided in the preface of this thesis.

The selected papers reprinted in this section are listed below.

Daryoush Habibi, DJH Lewis, DT Nguyen & Jason Pieloor, *'Analysis of an Access Node Multiplexer in a System Serving CBR and VBR Traffic'*, To appear in *'Computer Communications'*, 16(12), December 1993.

Daryoush Habibi, *'A Hidden Markov Model for Modelling VBR Video'*, Proceedings of the Seventh Australian Teletraffic Research Seminar, pages 181-190, Adelaide, November 1992.

Daryoush Habibi & DJH Lewis, *'Queues with Periodic Input and Output Rates'*, Proceedings of the Australian Broadband Switching & Services Symposium '93, pages 225-233, Wollongong, July 1993.

These articles have been removed for copyright or proprietary reasons.